REMARKS

Claims 1 and 3 are pending in the present application. Claim 3 has been canceled herein. Support for the amendment to claim 1 may be found in original claims 1 and 3. Additionally, claim 1 has been amended to positively recite, a "computationally filtered set." Support for the amendment may be found in the specification at page 26, line 26 through page 27, line 10 and page 28, lines 15-28.

Claim Rejection – 35 USC §101:
Claims 1 and 3 have been rejected under 35 U.S.C. 101 as having no specific or well-established utility.

The computational method of the present invention is directed at generating a secondary library of scaffold protein variants comprising: a) providing a primary library comprising a computationally filtered set of scaffold protein primary variant sequences; b) generating a list of primary variant positions in said primary library; c) combining a plurality of said primary variant positions to generate a secondary library of secondary sequences and d) synthesizing a plurality of said secondary sequences. The arguments set forth in the last Office Action regarding the computational method for screening variant sequences were merely supportive of the methodology's use. Therefore, there has been no change in the scope of the invention as claimed.

The Examiner asserts that because there is "no evidence of record or any line of reasoning that would support a conclusion that the secondary library was, as of the filing date, useful for any industrial or any pharmacological uses..." (Action, page 5). Applicants respectfully submit copies of publications in support of the method and that secondary libraries of the present invention find broad application and are useful. Applicants submit US Patent Nos. 6,627,186 and 6,514,729; Marshall, et. al, Rational Design and Engineering of Therapeutic Proteins, DDT ;Vol. 8, No. 5 March 2003; Luo, et. al, Development of a Cytokine Analog with Enhanced Stability Using Computational Ultrahigh throughput Screening, Protein Science (2002), 1:1218-1226; Filikov, et. al, Computational Stabilization of Human Growth Hormone, Protein Science (2002), 11:1452-1461; and DeGrado, William F., Proteins from Scratch, Science, 3 October 1997, Volume 278, pp. 80-81. These references provide support that the method of the present invention has specific and well-established utility.

The method of the present invention computationally generates ("screens") protein sequence libraries to select (produce) smaller libraries of protein sequences. This screening produces manageable libraries of proteins that may be synthesized and experimentally tested in an assay for a desired activity, for improved function and properties. The library can be computationally manipulated again, as the method of the present invention may be iterative, to create a new library, which then can be synthesized and experimentally tested, and so on.

The invention has two broad uses; first, the invention can be used to screen libraries based on known scaffold proteins. For example, computational screening for stability (or other properties) may be done on either the entire protein or some subset of residues. By using computational methods to utilize a threshold or cutoff to eliminate disfavored sequences (those not meeting a criteria for producing a desired characteristic, e.g., stability), the percentage of useful variants in a given variant set size increases, and the required experimental outlay is decreased.

The method of the present invention is also useful for screening random peptide libraries. Signaling pathways in cells often begin with an effector stimulus that leads to a phenotypically describable change in cellular physiology. Despite the key role intracellular signaling pathways play in disease pathogenesis, in most cases, little is understood about the signaling pathway aside from the initial stimulus and the ultimate cellular response. Historically, signal transduction has been analyzed by biochemistry or genetics. The biochemical approach dissects a pathway in a "stepping-stone" fashion: find a molecule that acts at, or is involved in, one end of the pathway, isolate assayable quantities and then try to determine the next molecule in the pathway, either upstream or downstream of the isolated one.

The genetic approach is classically a "shot in the dark": induce or derive mutants in a signaling pathway and map the locus by genetic crosses or complement the mutation with a cDNA library.

Limitations of biochemical approaches include a reliance on a significant amount of pre-existing knowledge about the constituents under study and the need to carry such studies out *in vitro*, post-mortem. The limitations of purely genetic approaches include the need to first derive and then characterize the pathway before proceeding with identifying and cloning the gene.

The literature is replete with examples of small peptides capable of modulating a wide variety of signaling pathways, which have been screened in *in vitro* assays for bioactivity.

Accordingly, generation of random or semi-random sequence libraries of proteins and peptides allows for the selection of proteins (including peptides, oligopeptides and polypeptides) with useful properties. The sequences in these experimental libraries can be randomized at specific sites only, or throughout the sequence. The number of sequences that can be searched in these libraries grows exponentially with the number of positions that are modified. Generally, only a small number sequences can be contained in an experimental library because of the physical constraints of laboratories (the size of the instruments, the cost of producing large numbers of biopolymers, etc.). These limits may be reached by selecting just a few amino acid positions to modify. Therefore, only a small sampling of sequences is possible to search for improved proteins or peptides in experimental sequence libraries. This lowers the chance of success and results in missing desirable variant candidates. Due to the randomness of changes in these techniques, most of the candidates in the library are not suitable, resulting in waste of much of the effort and resources used to produce the experimental library.

The present invention generates virtual libraries of protein sequences that are vastly larger than traditional experimental libraries. Many more candidate sequences may be screened computationally and those that meet design criteria, which favor stable and functional proteins, may be selected. An experimental library consisting of the favorable candidates found in the virtual library screening may then be generated, resulting in more efficient use of the experimental library and overcoming the limitations of random protein libraries.

By limiting the number of randomized positions and the number of possibilities at these positions, the number of wasted sequences produced in the experimental library is reduced, thereby increasing the likelihood of finding sequences with the desired properties.

Additionally, by computationally screening large libraries, greater diversity of protein sequences may be screened (i.e. a larger sampling of sequence space), leading to greater improvements in protein function. Furthermore, fewer variant sequences need to be experimentally tested (physically generated) to screen a given library, reducing the cost and difficulty of protein engineering. By using computational methods to screen protein libraries, speed and efficiency are combined with the ability of experimental library screening to create new activities in proteins for which appropriate computational models and structure-function relationships are unclear.

The method of the present invention provides for the biasing of libraries in any number of ways (filtered set), allowing the generation of secondary libraries having desired characteristics, e.g., improved function or stability. For example, domains, subsets of residues, active or binding sites, surface residues, etc. may be selected, thereby increasing the diversity of sequences generated by the method of the present invention.

In addition, Applicants submit that there is adequate support at the Specification at page 5, line 13 through page 8, line 24, for both specific utility and well-established utility.

In light of the above-referenced publications, the support in the specification and Applicants discussion above, Applicants submit that the method and resulting secondary libraries have both a specific and well-established utility.

Claim Rejection – 35 USC §112, First Paragraph:

Claims 1 and 3 are rejected under 35 USC §112, first paragraph because the specification while enabling for the enzymes protein design using specific program design, does not reasonably provide enablement for any type of secondary library of scaffold protein variants or sequences. Claim 3 has been canceled herein making rejection moot.

Applicability of the method is not speculative or unexplored, as the Action suggests. The Action states that the accuracy of the statements in the specification and claims must be sufficiently supported by well-established chemical principles or by sufficient number of examples. The present application is founded on well-established principles of chemistry. The present invention utilizes well-established protein design methods, which are founded on basic principles of chemistry, e.g., utilizes structural and biophysical knowledge of proteins. See for example, Specification at page 1, lines 11-25.

The Action cites the "high unpredictability of the newly emerging biolibrary art…" (page 9) page 1, lines 13-25 and page 6, line3 to page 7, line 3. Applicants respectfully submit that the cited references of "biolibraries" are distinguishable from the present invention because the cited sections discuss non-rational methods in current use for screening libraries of mutants, and which are highly unpredictable.

The present invention is clearly distinguishable from the "directed molecular evolution" mutagenesis techniques referenced in the Action because as stated above, it is a rational design technique, not a randomly generated library, as is done with directed molecular evolution.

The present method may further be distinguished from the referenced biochemical and genetic methods, again because it generates secondary library members by in a rational way. The present invention allows the screening of a substantially greater number of variants because it utilizes structural and biophysical knowledge of the target proteins (well-established principles of chemistry.) See for example, Specification at page 1, lines 11-25.

As stated above, the present invention provides significantly diverse results because is a rational method. The directed molecular evolution techniques have limited diversity because their libraries are generated using multiple sequences whose fragments are mixed together and re-assembled based on pre-identified cross-over points (sections of sequence homology among the initial sequences). Thus, diversity is limited mixtures of the initial sequences chosen.

The enablement requirement refers to the requirement of 35 USC 112, first paragraph that the specification describe how to make and how to use the invention. One skilled in the art must be enabled to make and use is that defined by the claim(s) of the particular application or patent. Applicants submit that in light of the distinction between rational and non-rational protein design methods and the above-arguments, the method of the present invention is enabled by the disclosure.

In conclusion, Applicants submit that the Specification taken in conjunction with the state of the art at the time the invention was filed and the evidence in support of the broad applicability of the method fully enables a person skilled in the art to practice the invention without undue experimentation. Applicants respectfully request reconsideration and withdrawal of the rejection of the claim.

Claim Rejection – 35 USC §112, Second Paragraph:

Claims 1 and 3 are rejected under 35 USC 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention for reasons of record. Claim 3 has been canceled herein, so only claim 1 will be addressed below.

Applicants submit that the filtered set is not necessarily obtained by a rank ordered list or by a scoring function. In some cases the user of the method will use selection criteria that may or may not be a rank ordered list of sequences, as that is only one embodiment. Other criteria may be used to bias the list. In response, Applicants disagree with the limitation of a rank ordered list. See Specification beginning on page 26, line 26 through page 27, line10.

In light of the forgoing argument Applicants respectfully request reconsideration and withdrawal of the rejection of the claim.

Claim Rejection – 35 USC §102:

Claims 1 and 3 are rejected under 35USC 102 as being anticipated by Fechteler et al.

The Fechteler reference is a homology modeling paper about predicting protein structure in regions of insertions and deletions. This reference is directed to designing a protein model by homology and predicting what structure the sequence would adopt. The reference does not retain the sequence information because they are focused on the backbone structure.

"A claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." *Verdegaal Bros. v. Union Oil Co. of California*, 814 F.2d 628, 631, 2 USPQ2d 1051, 1053 (Fed. Cir. 1987). "The identical invention must be shown in as complete detail as is contained in the ... claim." *Richardson v. Suzuki Motor Co.*, 868 F.2d 1226, 1236, 9 USPQ2d 1913, 1920 (Fed. Cir. 1989)."

The present invention may be distinguished from the cited reference because there is no suggestion or teaching of synthesizing variants of a secondary library. The present invention is not directed to predicting theoretical 3D structures based solely on homology models and does not rely on insertion and deletion regions. Therefore, the claim as amended, is not anticipated by the cited reference because each and every element as set forth in the claim is not found, either expressly or inherently described, in the reference. In light of the foregoing, Applicants respectfully request reconsideration and withdrawal of the claim rejections.

DOUBLE PATENTING:

Claims 1 and 3 are rejected under the judicially created doctrine of obviousness-type double patenting as being unpatentable over claims 1-2 of US 6, 403, 312.

Applicants respectfully submit that a terminal disclaimer is being filed herewith for the Examiner's review, thereby making the double patenting rejection moot.

The Applicants submit that in light of the above-amendment and argument submission of a terminal disclaimer, the claims are now in condition for allowance and an early notification of such is respectfully solicited.

Please direct any calls in connection with this application to the undersigned at (626) 737-8019.

Dated: December 19, 2003
Xencor
111 W. Lemon Avenue
Monrovia, California 91016
Phone: (626) 737-8019
Fax: (626) 256-3760

Respectfully submitted,

By:
Joyce L. Morrison, Reg. No. 31,902

Computational stabilization of human growth hormone

ANTON V. FILIKOV, ROBERT J. HAYES, PEIZHI LUO, DIANE M. STARK,
CHERYL CHAN, ANIRBAN KUNDU, AND BASSIL I. DAHIYAT
Xencor, Inc., Monrovia, California 91016, USA

(RECEIVED August 22, 2001; FINAL REVISION January 14, 2002; ACCEPTED February 18, 2002)

Abstract

Recombinant human growth hormone (hGH) is used worldwide for the treatment of pediatric hypopituitary dwarfism and in children suffering from low levels of hGH. It has limited stability in solution, and because of poor oral absorption, is administered by injection, typically several times a week. Development has therefore focused on more stable or sustained-release formulations and alternatives to injectable delivery that would increase bioavailability and make it easier for patients to use. We redesigned hGH computationally to improve its thermostability. A more stable variant of hGH could have improved pharmacokinetics or enhanced shelf-life, or be more amenable to use in alternate delivery systems and formulations. The computational design was performed using a previously developed combinatorial optimization algorithm based on the dead-end elimination theorem. The algorithm uses an empirical free energy function for scoring designed sequences. This function was augmented with a term that accounts for the loss of backbone and side-chain conformational entropy. The weighting factors for this term, the electrostatic interaction term, and the polar hydrogen burial term were optimized by minimizing the number of mutations designed by the algorithm relative to wild-type. Forty-five residues in the core of the protein were selected for optimization with the modified potential function. The proteins designed using the developed scoring function contained six to 10 mutations, showed enhancement in the melting temperature of up to 16°C, and were biologically active in cell proliferation studies. These results show the utility of our free energy function in automated protein design.

Keywords: Protein design; free energy; entropy; human growth hormone; thermostability

Human growth hormone (hGH) is a polypeptide hormone that is synthesized by the somatotropic cells of the anterior pituitary. It plays an important rôle in somatic growth through its effects on the metabolism of proteins, carbohydrates, and lipids. hGH is currently used for the treatment of pediatric hypopituitary dwarfism and in children suffering from low levels of hGH (Hindmarsh and Brook 1987). It is believed that hGH functions by direct action on bone and soft tissue to cause uniform growth and by indirect stimulation of insulin-like growth factor-1 (Pearlman and Bewley 1993).

The most prevalent form of pituitary hGH is a single-chain polypeptide containing 191 amino acids, internally cross-linked by two disulfide bonds. The molecular mass is ~22 kD, with pI near 5.3. Approximately 55% of the polypeptide backbone exists in a right-handed α-helical conformation. The hormone is a four-helix bundle showing an up-up-down-down topology. Activation of transmembrane receptors for hGH (hGHbp) occurs when dimerization of receptor chains is triggered by binding of hGH to a ligand-binding domain on the receptor. The crystal structure of the wild-type hGH in 1:2 complex with its receptor was determined to 2.8-Å resolution (de Vos et al. 1992). There are other crystal structures of the protein, its mutants, and complexes available in the literature and the Protein Data Bank (PDB; Sundstrom et al. 1996; Atwell et al. 1997; Clackson et al. 1998).

Met-hGH and hGH are produced recombinantly and are available worldwide for clinical use. Both forms have therapeutic activity that is equivalent to the pituitary-derived material (Jorgensen 1987). Because hGH is a protein, it is not

Reprint requests to: Bassil I. Dahiyat, Xencor, Inc., 111 W. Lemon Avenue, Monrovia, CA 91016; e-mail: baz@xencor.com; fax: (626) 305-0350.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.3500102.

absorbed orally to any significant extent (Moore et al. 1986) and must be administered by injection. It is typically given subcutaneously or intramuscularly several times a week over an extensive period. It has limited stability in solution (for ~2 weeks at 2°C to 8°C) and is commonly stored in freeze-dried form. Development has therefore focused on more stable or sustained-release formulations and alternatives to injectable delivery that would increase bioavailability and make it easier for patients to use. In this study, we have redesigned hGH computationally to improve its thermostability. A more thermostable variant of hGH could have improved utilization time or a longer shelf-life, which would translate into decreased costs for the manufacturer and added convenience and compliance for patients. Thermostability, described in terms of T_m, or the denaturation temperature of unfolding, has been used to predict the best long-term storage conditions for protein pharmaceuticals (Schrier et al. 1993; Remmele et al. 1998). A more stable variant of hGH could also have improved pharmacokinetics or be more amenable to use in alternative delivery systems and formulations.

There are two components required for computational design: (1) accurate scoring functions to rank sequences and (2) high-speed optimization methods to rapidly find the best sequences from the enormous combinatorial search space (Dahiyat 1999). We use our Protein Design Automation (PDA) method (Dahiyat and Mayo 1996, 1997a; Dahiyat et al. 1997), which incorporates the dead-end elimination (DEE) algorithm (Desmet et. al 1992; Goldstein 1994). Using a rotamer description of the side-chains, an optimal sequence for a backbone can be found by screening all possible sequences of rotamers, in which each backbone position can be occupied by each amino acid in all possible rotameric states.

The scoring functions used for protein design were recently reviewed by Gordon et al. (1999). Although nonenergy terms such as secondary structure propensities can be used, the most successful designs use energy functions based on molecular mechanics force field terms (van der Waals, hydrogen bonding, electrostatics, bond and angle energy), that is, potential energy terms, or their combinations with free energy terms such as solvation (Dahiyat and Mayo 1996) or entropy (Hellinga and Richards 1994; Dahiyat and Mayo 1996; Kono et al. 1998). Here we use a previously developed scoring function that includes potential energy terms (van der Waals, hydrogen bonding, electrostatics) and solvation terms (polar hydrogen burial and nonpolar exposure penalties, nonpolar burial energy) augmented with a term that accounts for the loss of backbone and side-chain conformational entropy. Before side-chain selection, residues are identified as core, surface, or boundary using the RESCLASS residue classification program (Dahiyat and Mayo 1997a).

Combining potential energy and free energy terms to estimate the free energy of folding or binding assumes both additivity and proportionality of potential energy and free energy terms. This necessarily raises the question of proper weighting factors for the terms. In the simplest treatment, the weighting factors are assumed to be equal to one (Kono et al. 1998). Alternatively, the factors can be derived by regression to experimental free energy data (Dahiyat and Mayo 1996; Filikov and James 1998; Filikov et al. 2000). Here we use a different approach: The weighting factors are optimized by minimizing the number of mutations designed by the algorithm.

The loss of entropy on formation of the folded protein is believed to be the principal force opposing folding (Stites and Pranata 1995). Therefore, inclusion of side-chain and main-chain entropy terms into the scoring function with proper weighting factors could improve scoring of designed sequences. A side-chain entropy term has been incorporated into protein design energy functions previously (Hellinga and Richards 1994; Kono et al. 1998). The change in side-chain entropy on folding can be modeled as the change in the number of rotatable bonds, assuming that conformational freedom is completely restricted in the folded state (Hellinga and Richards 1994). An empirical approach is based on the entropy of fusion of small organic compounds (Sternberg and Chickos 1994). Alternatively, the change in entropy can be derived from the distribution of side-chain rotamers in crystal structures (Pickett and Sternberg 1993) or in Monte Carlo simulations (Creamer and Rose 1992; Creamer 2000). These and other methods of estimating conformational entropy have been described recently (Creamer 2001) and were shown to correlate extremely well, despite different methods of derivation. In this study, we use both side-chain and backbone entropy terms based on scales introduced by Pickett and Sternberg (1993) and by Stites and Pranata (1995), respectively (Table 1).

Results

The scoring function used in this work is a sum of the following terms: van der Waals interaction, hydrogen bond potential, distance-dependent Coulombic electrostatics, polar hydrogen burial penalty, nonpolar burial energy, nonpolar exposure penalty, and entropy. A detailed description of all the terms, except the entropy term, is given elsewhere (Dahiyat and Mayo 1997a,b). Here we optimize the weighting factors for the entropy (λ_S) and polar hydrogen burial penalty (ΔG_H) terms, and the dielectric constant (ε) for the electrostatic term. In the following sections, we describe independent optimization of each parameter, beginning with the entropy term. Although simultaneous optimization of the three weighting factors is a possibility, such an approach is often problematic because of correlations between parameters. Here, by focusing on different sets of residue classes

Table 1. *Values of $T\Delta S$ (kcal/mol) for the amino acids at 20°C*

Amino acid	Side-chain $T\Delta S$[a]	Backbone $T\Delta S$[b]	Total $T\Delta S$[c]
Ala	0.0	−0.71	−1.21
Arg	−2.03	−0.51	−3.44
Asn	−1.57	−0.18	−3.31
Asp	−1.25	−0.29	−2.88
Cys	−0.55	−0.29	−2.18
Gln	−2.11	−0.48	−3.55
Glu	−1.81	−0.64	−3.09
Gly	0.0	0.0	−1.92
His	−0.96	−0.21	−2.67
Ile	−0.89	−0.59	−2.22
Leu	−0.78	−0.55	−2.15
Lys	−1.94	−0.42	−3.44
Met	−1.61	−0.51	−3.02
Phe	−0.58	−0.31	−2.19
Pro	0.0	−0.82	−1.10
Ser	−1.71	−0.28	−3.35
Thr	−1.63	−0.29	−3.26
Trp	−0.97	−0.44	−2.45
Tyr	−0.98	−0.32	−2.58
Val	−0.51	−0.57	−1.86

[a] Taken from Pickett and Sternberg (1993).

[b] Taken from Sites and Pranata (1995).

[c] Obtained by summing up the side-chain scale and the backbone scale, corrected for the glycine backbone entropy loss (−1.92 kcal/mole) taken from D'Aquino et al. (1996); $T\Delta S = T\Delta S_{\text{side-chain}} + (-1.92 - T\Delta S_{\text{backbone}})$.

that are predominantly dependent on one of the parameters, we are able to optimize each parameter independently, thus minimizing the possibility of spurious results. Furthermore, simultaneous optimization is significantly more computationally intensive because it requires that a much more extensive set of calculations be performed and analyzed.

Weighting factor for the entropy term

We optimized the weighting factor for the entropy term by minimizing the number of mutations designed by the algorithm. This approach is based on the assumption that the wild-type sequence is reasonably close to the global energy minimum (GEM) in the sequence space of a particular fold (Kuhlman and Baker 2000), and by minimizing the distance from the wild-type sequence, we minimize the distance from the GEM sequence. The wild-type sequence often is not the GEM sequence, because stabilizing mutations for numerous proteins are known. For example, in this work we find two sequences that are considerably more thermostable than the wild type. Without knowing the GEM sequence, however, a reasonable option is to use the wild-type sequence as a target for optimization of the algorithm parameters. To derive and validate a broadly applicable parameter set, a number of different proteins should be used to optimize parameter values. Here we derive a parameter set based only on hGH, which will be tested more extensively in future work.

We selected 45 residues buried in the core of hGH for entropy calculations. Residue classification with RESCLASS gives 71 core residues for hGH (PDB structure 3HHR). To make the calculations faster and to focus the optimization on residues for which the entropy term is isolated from the electrostatic and polar hydrogen energies, we reduced this list to 45 positions by eliminating residues involved in hydrogen bonds and residues with significant exposure to the solvent. Several rounds of design were performed with PDA using different weighting factors for the entropy term in the range of one to four. For each round of design, the DEE algorithm was run to completion; that is, the global energy minimum sequence (GEMS) was identified. The number of mutations contained in the GEMS strongly depends on the entropy term weighting factor and has a clear minimum centered at 2.2 (Table 2; Fig. 1). At smaller entropy weighting factors (<1.7), the GEMS tends to contain a lot of methionine residues (methionine is very flexible and can fill cavities of a wide variety of shapes). Simultaneously, the loss of total entropy on folding (*$T\Delta S$*) for methionine is very high (−3.02 kcal/mole). Frequent appearance of methionines in the GEMS is the most obvious consequence of neglecting the entropy term in the scoring function. At higher entropy weighting factors (>3), the GEMS tends to have larger residues mutated to smaller ones, that is, to less entropically rich residues: Ile→Val, Leu→Ala, and Met→Ala (see Table 2). The optimal value for the entropy weighting factor is in the range of 1.7 to 2.7, as can be seen from Figure 1.

Weighting factor for the electrostatic term

The same approach was used for optimization of the weighting factor for the electrostatic term. However, the set of core residues used for the entropy calculations cannot be used to optimize the electrostatics term, because there are few polar residues in the core. For these calculations, we selected a set of 28 boundary residues. These were obtained by running the RESCLASS algorithm, which gives 41 boundary residues for hGH, and eliminating the residues within 5 Å of the receptor and Gly104, because it has unusual ϕ and ψ angles. The result is a set of 28 residues that are predominantly buried: The solvent accessible fraction of the residue surface is 32.7% on average. Therefore, we assume that for these residues all the entropy of the unfolded state is lost on folding and treat them no differently from core residues in this respect.

Ten rounds of design were performed using different values of the dielectric constant for the electrostatic term in the range of 5R to 40R, where R is the interatomic distance (varying ε is equivalent to varying the weighting factor). The weighting factor for the entropy term was set to 2.3, the midpoint of the optimal values found previously (Fig. 1). For each round of design, the DEE algorithm was run to

Table 2. *Global energy minimum sequences found by PDA for different entropy weighting factors (λ_s)*

Position[a]	Wild type	λ_s							
		1	1.4	1.7	2	2.3	2.7	3	4
6	Leu	—[b]	—	—	—	—	—	—	—
10	Phe	—	—	—	—	—	—	—	—
13	Ala	Val	Val	.Val	Val	Val	Val	Val	Val
17	Ala	—	—	—	—	—	—	—	—
20	Leu	Met	Met	—	—	—	—	—	Ile
24	Ala	—	—	—	—	—	—	—	—
27	Thr	Val	Val	Val	Val	Val	Val	Val	Val
28	Tyr	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe
31	Phe	—	—	—	—	—	—	—	—
36	Ile	—	—	—	—	—	—	—	—
44	Phe	—	—	—	—	—	—	—	—
54	Phe	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr
55	Ser	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
58	Ile	—	—	—	—	—	—	Val	Val
73	Leu	—	—	—	—	—	—	Ala	Ala
75	Leu	—	—	—	—	—	—	—	—
76	Leu	—	—	—	—	—	—	—	—
78	Ile	—	—	—	—	—	—	—	—
79	Ser	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
80	Leu	—	—	—	—	—	—	—	—
81	Leu	—	—	—	—	—	—	—	—
82	Leu	—	—	—	—	—	—	—	—
83	Ile	—	—	—	—	—	—	—	Val
85	Ser	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
90	Val	Ile	Ile	—	—	—	—	—	—
93	Leu	—	—	—	—	—	—	—	—
96	Val	—	—	—	—	—	—	—	—
97	Phe	—	—	—	—	—	—	—	—
105	Ala	—	—	—	—	—	—	—	—
110	Val	Met	Met	—	—	—	—	—	—
114	Leu	Met	Met	Met	Met	Met	Met	—	Phe
117	leu	Met	Met	—	—	—	—	—	—
11	Ile	Val	Val	—	—	—	—	—	—
124	Leu	—	—	—	—	—	—	—	—
157	Leu	—	—	—	—	—	—	—	—
161	Gly	Met	Met	Met	Met	Met	Met	Met	Met
162	Leu	—	—	—	—	—	—	—	—
163	Leu	—	—	—	—	—	—	—	—
166	Phe	Leu	Leu	Leu	Leu	Leu	Leu	Met	Leu
170	Met	—	—	—	—	—	—	Leu	Ala
173	Val	—	—	—	—	—	—	—	—
176	Phe	—	—	—	—	—	—	—	—
177	Leu	—	—	—	—	—	—	—	—
180	Val	—	—	—	—	—	—	—	—
184	Ser	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala

[a] The residues are numbered as in 3HHR file from Brookhaven Protein Databank.

[b] "—" indicates the wild-type residue.

completion. The number of mutations contained in the GEMS is plotted versus the dielectric constant in Figure 2. As can be seen, the curve has a distinct minimum at ε/R = 10.3 ± 0.9.

At low dielectric constants, PDA tends to place charged or polar residues; at higher constants, these mutate to wild-type or non–wild-type uncharged or apolar residues (see Table 3). Examples of this trend include positions 34, 35, 71, 84, and 157. On the other hand, at high constants, some charged positions, including the wild-type ones, mutate to apolar amino acids. Examples of this trend include positions 74 and 118. Superposition of these two trends results in a curve with a minimum at ε/R = 10.3 ± 0.9.

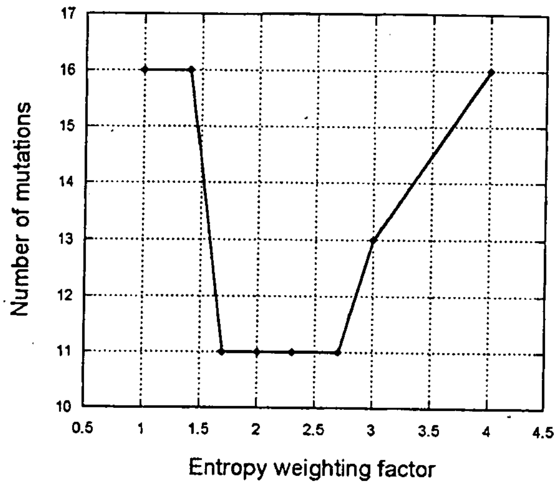


Fig. 1. Number of mutations versus weighting factor for the entropy term.

Polar hydrogen burial penalty term

To optimize the polar hydrogen burial penalty term, we ran 11 rounds of design with values of the penalty from 0 to 3 kcal/mole. The dielectric constant was set to 10.3R, the optimal value obtained previously (Fig. 2). The entropy term weighting factor and other parameters were the same as in the optimization of the dielectric constant, as were the

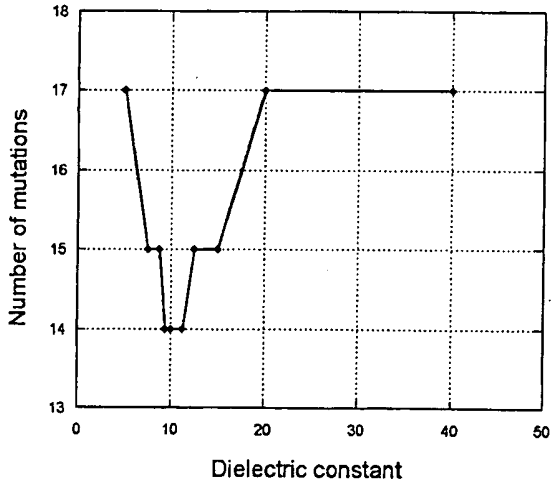


Fig. 2. Number of mutations versus dielectric constant.

Table 3. *Global energy minimum sequences found by PDA for different dielectric constants (ϵ)*

	Wild	ϵ/R, R = interatomic distance									
Position[a]	type	40	20	15	12.5	11.25	10	9.37	8.75	7.5	5
6	Leu	—[b]	—	—	—	—	—	—	—	—	—
14	Met	Phe	Phe	Leu	Leu	Leu	Leu	Leu	Leu	Leu	Leu
26	Asp	—	—	—	—	—	—	—	—	—	—
30	Glu	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp
32	Glu	—	—	—	—	—	—	—	—	—	—
34	Ala	—	—	—	—	—	—	—	Ser	Ser	Hsp
35	Tyr	—	—	—	—	—	—	—	—	—	Asp
40	Gln	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp
50	Thr	Met	Met	Met	Met	Met	Met	Met	Met	Met	Met
56	Glu	—	—	—	—	—	—	—	—	—	—
57	Ser	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr
59	Pro	Val	Val	Val	Val	Val	Val	Val	Val	Val	Val
66	Glu	—	—	—	—	—	—	—	—	—	—
71	Ser	Thr	Thr	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp
74	Glu	Phe	Phe	—	—	—	—	—	—	—	—
84	Gln	Ile	Ile	Ile	Ile	Ile	Ile	Ile	Ile	Lys	Lys
92	Phe	—	—	—	—	—	—	—	—	—	—
107	Asp	Ala	Ala	Ala	Ala	—	—	—	—	—	—
109	Asn	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe
113	Leu	—	—	—	—	—	—	—	—	—	—
118	Glu	Phe	Phe	—	—	—	—	—	—	—	—
125	Met	—	—	—	—	—	—	—	—	—	—
130	Asp	His	His	His	His	His	His	His	His	His	His
139	Phe	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
143	Tyr	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
157	Leu	—	—	—	—	—	—	—	—	—	Hsp
158	Lys	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe
183	Arg	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp

[a] The residues are numbered as in 3HHR file from Brookhaven Protein Databank.
[b] "—" indicates the wild-type residue.

residues selected for design (28 boundary residues). The number of mutations contained in the GEMS is plotted versus the polar hydrogen burial penalty in Figure 3. The optimal value for the penalty is 1.6 ± 0.6 kcal/mole.

At low values of the penalty, charged or polar residues appear at some positions and become apolar or less polar as the penalty increases (see Table 4). This is the case for positions 40, 57, 71, 84, 139, 143, and 157. At high values of the penalty ($\Delta G_H \geq 2.5$), wild-type Glu at position 74 mutates to Phe; that is, a charged residue mutates to an apolar one. Superposition of these two trends results in a curve with a minimum at 1.6 ± 0.6 kcal/mole.

Redesign of the core of hGH

To enhance the thermostability of hGH, we used PDA to computationally redesign 45 residues in the core of the protein (the same set that was used in the entropy weighting factor optimization). We used the parameters optimized as described above: entropy term with weighting factor of 2.3, penalty for polar hydrogen burial of 1.6 kcal/mole, and dielectric constant of 10.3R. The surface-based nonpolar exposure penalty and nonpolar burial benefit were set to 0.048 kcal/mole/Å^2. The calculation resulted in 11 mutations

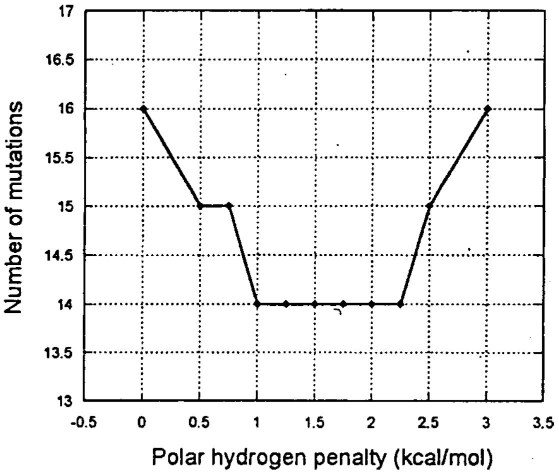


Fig. 3. Number of mutations versus polar hydrogen burial penalty.

Table 4. *Global energy minimum sequences found by PDA for different polar hydrogen burial penalties (ΔG_H)*

		ΔG_H (kcal/mole)										
Position[a]	Wild type	0	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	3
6	Leu	—[b]	—	—	—	—	—	—	—	—	—	—
14	Met	Leu	Leu	Leu	Leu	Leu	Leu	Leu	Leu	Leu	Leu	Leu
26	Asp	—	—	—	—	—	—	—	—	—	—	—
30	Glu	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp	Trp
32	Glu	—	—	—	—	—	—	—	—	—	—	—
34	Ala	—	—	—	—	—	—	—	—	—	—	—
35	Tyr	—	—	—	—	—	—	—	—	—	—	—
40	Gln	Arg	Arg	Arg	Arg	Arg	Arg	Arg	Arg	Arg	Arg	Arg
50	Thr	Phe	Phe	Phe	Met	Met	Met	Met	Met	Met	Met	Met
56	Glu	—	—	—	—	—	—	—	—	—	—	—
57	Ser	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	Ala	Ala	Ala
59	Pro	Val	Val	Val	Val	Val	Val	Val	Val	Val	Val	Val
66	Glu	—	—	—	—	—	—	—	—	—	—	—
71	Ser	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Thr	Thr
74	Glu	—	—	—	—	—	—	—	—	—	Phe	Phe
84	Gln	Arg	Lys	Lys	Lys	Lys	Lys	Ile	Ile	Ile	Ile	Ile
92	Phe	—	—	—	—	—	—	—	—	—	—	—
107	Asp	—	—	—	—	—	—	—	—	—	—	—
109	Asn	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Tyr	Tyr
113	Leu	—	—	—	—	—	—	—	—	—	—	—
118	Glu	Leu	—	—	—	—	—	—	—	—	—	—
125	Met	—	—	—	—	—	—	—	—	—	—	Val
130	Asp	His	His	His	His	His	His	His	His	His	His	His
139	Phe	His	His	His	His	His	His	His	Ala	Ala	Ala	Ala
143	Tyr	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
157	Leu	Arg	Arg	Arg	—	—	—	—	—	—	—	—
158	Lys	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe	Phe
183	Arg	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp	Hsp

[a] The residues are numbered as in 3HHR file from Brookhaven Protein Databank.
[b] "—" indicates the wild-type residue.

(Table 5). We selected sequences for experimental testing by ranking the mutations according to their contribution to lowering the energy of the wild-type sequence. The six highest-ranking mutations were selected for the CORE1 sequence; two others were added to obtain CORE2; and another two were added to obtain CORE3 (Table 5). The model structure and 10 mutations of the CORE3 protein are shown in Figure 4.

Thermal stability

CORE1, CORE2, and CORE3 proteins and wild-type hGH were expressed and isolated as described in Materials and Methods. The far-ultraviolet circular dichroism spectra for the proteins were nearly identical to each other and to the wild-type protein, indicating highly similar secondary structure and tertiary folds (data not shown). Thermal denaturation was monitored at 222 nm for wild-type hGH, CORE1, and CORE2 (Fig. 5; data was not obtained for CORE3).

The melting temperatures (T_ms) were estimated graphically by finding the midpoints on the transition region of the melting curve. Because the T_ms for the mutants are close to 100°C and the ends of the transition regions of the curves are beyond the experimental range, only the lower bounds of the T_ms can be estimated. This gives the following values: wild-type $T_m = 82°C$, CORE1 $T_m \geq 98°C$, and CORE2 $T_m \geq 95°C$. The designed proteins thus showed enhancements of 13°C to 16°C. It should be noted that thermal melting was not reversible as measured; therefore, the T_m values given here are not rigorous thermodynamic parameters. However, these values are indicative of the improved thermostability of the designed proteins.

Biological activity

The biological activity of CORE1, CORE2, and CORE3 proteins was determined in vitro by quantitating cell proliferation as a function of protein concentration. Figure 6 shows the dose-response curves of CORE1, CORE2, CORE3, and wild-type hGH in a representative assay. EC_{50}

Table 5. *Global energy minimum sequence found by PDA using optimized parameters for a core design of hGH[a] (Design) and experimentally tested sequences (CORE1, CORE2, CORE3)*

Position[b]	Wild type	Design	CORE1	CORE2	CORE3
6	Leu	—[c]	—	—	—
10	Phe	—	—	—	—
13	Ala	Val	Val	Val	Val
17	Ala	—	—	—	—
20	Leu	—	—	—	—
24	Ala	—	—	—	—
27	Thr	Val	Val	Val	Val
28	Tyr	Phe	—	—	Phe
31	Phe	—	—	—	—
36	Ile	—	—	—	—
44	Phe	—	—	—	—
54	Phe	Tyr	—	—	Tyr
55	Ser	Ala	—	Ala	Ala
58	Ile	—	—	—	—
73	Leu	—	—	—	—
75	Leu	—	—	—	—
76	Leu	—	—	—	—
78	Ile	—	—	—	—
79	Ser	Ala	Ala	Ala	Ala
80	Leu	—	—	—	—
81	Leu	—	—	—	—
82	Leu	—	—	—	—
83	Ile	—	—	—	—
85	Ser	Ala	—	Ala	Ala
90	Val	Ile	Ile	Ile	Ile
93	Leu	—	—	—	—
96	Val	—	—	—	—
97	Phe	—	—	—	—
105	Ala	—	—	—	—
110	Val	—	—	—	—
114	Leu	Met	—	—	—
117	Leu	—	—	—	—
121	Ile	—	—	—	—
124	Leu	—	—	—	—
157	Leu	—	—	—	—
161	Gly	Met	Met	Met	Met
162	Leu	—	—	—	—
163	Leu	—	—	—	—
166	Phe	—	—	—	—
170	Met	—	—	—	—
173	Val	—	—	—	—
176	Phe	—	—	—	—
177	Leu	—	—	—	—
180	Val	—	—	—	—
184	Ser	Ala	Ala	Ala	Ala

[a] Protein Data Bank structure 3HHR.
[b] The residues are numbered as in 3HHR file from Brookhaven Protein Databank.
[c] "—" indicates the wild-type residue.

values were determined by nonlinear least-squares fit of sigmoidal parts of the averaged curves to a four-parameter sigmoidal equation as described in Materials and Methods. The designed proteins showed comparable activity to wild-type hGH (Table 6).

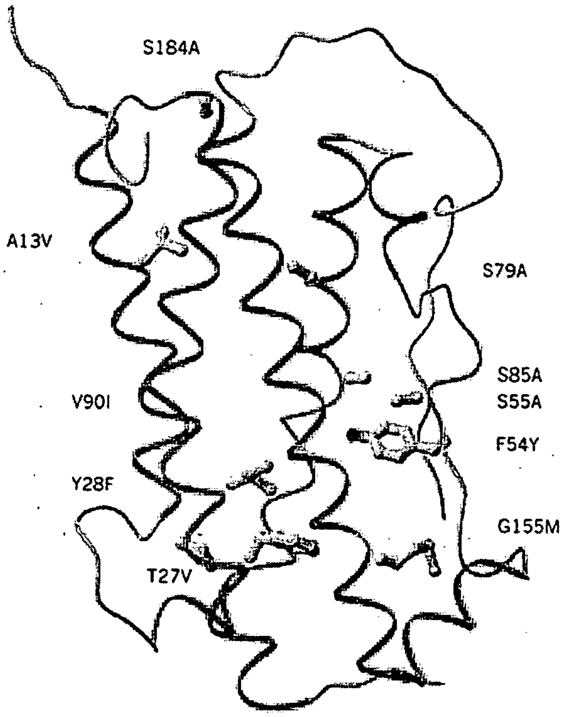


Fig. 4. Structure of CORE3 protein (a model generated by PDA, our computational design method). The 10 mutations are shown in ball-and-stick representation (hydrogen atoms are not shown).

Discussion

This study has two purposes: (1) improving our sequence energy scoring function by both adding an entropy term and optimizing the relative weights of the energy terms, and (2) improving the thermostability of hGH. We designed only

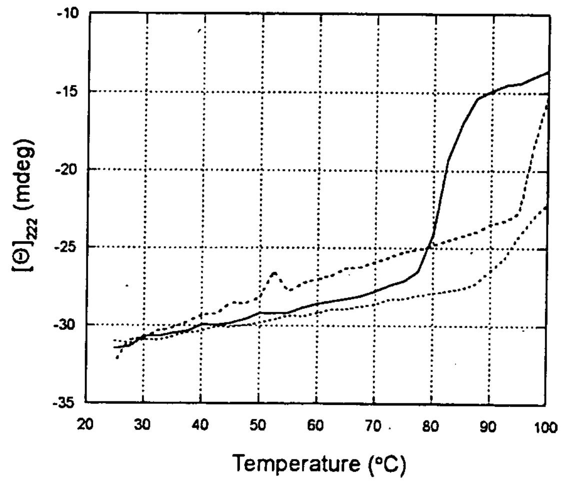


Fig. 5. Thermal denaturation monitored by circular dichroism at 222 nm for wild-type hGH (solid line), CORE1 (dashed line), and CORE2 (dotted line).

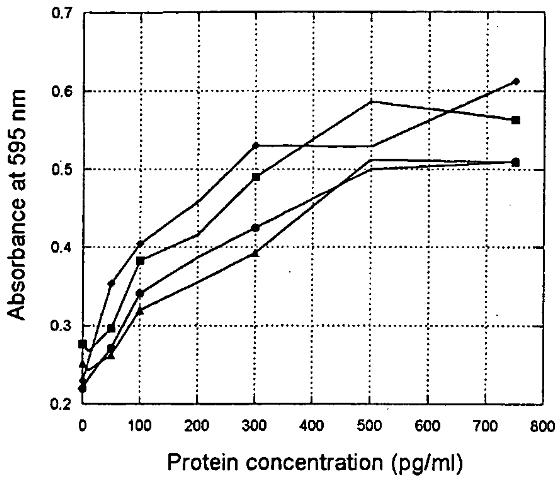


Fig. 6. Proliferation of BAF/B03 cells expressing the hGH receptor in response to wild-type hGH (diamonds), CORE1 (squares), CORE2 (triangles), and CORE3 (circles). Each point represents the average of three replicates.

the core residues of the protein, rather than the surface-exposed residues, to reduce the probability of an immunogenic response to the mutated protein. Designing only core residues simplified implementation of the entropy penalty. Core residues can be modeled simply as losing all entropy relative to the free side-chain, whereas boundary and surface residues require correction factors, such as scaling based on accessible surface area (Abagyan and Totrov 1994), to account for remaining conformational flexibility in the folded state. Designing only core residues also simplified electrostatic modeling. Optimization of the weighting factor for electrostatic energy showed that a large distance-dependent dielectric (ε/R ~10) was necessary to reduce the magnitude of the electrostatic energy and mitigate inaccuracies in the charge model, a weakness of all force fields. Because no charged residues were in the core design, the inaccuracies of force field approaches for charge-charge interactions were eliminated from our hGH variants.

In this work, we optimize the weighting factors by minimizing the number of mutations designed by the algorithm. That is, we assume that the wild-type sequence approximately corresponds to the global energy minimum in the sequence space. This is more correct for highly stable proteins, which were optimized for stability by nature. Therefore, further development of this idea should include calculations on a test set of several highly stable proteins with known high-resolution X-ray structures. An alternative approach to optimize the weighting factors is the use of mutagenesis data to correlate mutant stability with the energy function predictions. Of particular interest, of course, is testing the stability of the sequences designed by the algorithm. Unfortunately, this is a very time-consuming approach.

Table 6. *Biological activity of the designed proteins*

Protein	EC$_{50}$ (pg/mL)
Wildtype hGH	220 ± 20
CORE1	320 ± 30
CORE2	260 ± 50
CORE3	230 ± 50

The increased stability seen with CORE1 and CORE2 results from improved van der Waals packing interactions and increased burial of hydrophobic groups (A13V, T27V, V90I, G161M) and from replacement of unsatisfied hydrogen bond donors or acceptors with hydrophobic residues (T27V, S55A, S79A, S85A, S184A). It should be noted that our design resulted in the replacement of one threonine and four serines, residues that do not seem to form hydrogen bonds in the native protein. Although the role of these T→A and S→A mutations has not been determined individually, the considerable improvements in the T$_m$s obtained indicate that these mutations are beneficial for stability.

We obtained highly stabilized variants of hGH, a result of considerable practical interest and potential clinical significance. An equipotent, but more robust, hGH molecule could have improved pharmacokinetics or better storage properties or be more amenable to use in alternative delivery systems and formulations, thus providing added convenience and improved patient compliance. Also, the large increase in T$_m$ (≥16°C) shows the utility of our optimized energy function in automated protein design.

Materials and methods

Entropy term

We used the side-chain entropy scale taken from Pickett and Sternberg (1993) and the backbone entropy scale from Stites and Pranata (1995). Both scales were derived by analyzing the distribution of side-chain rotamers and backbone angles in crystal structures. We assume that all the entropy is lost on folding, because all the designed residues in the current work are mostly buried in the core of the protein. Therefore, our entropy scale is obtained by summing up the side-chain entropy scale and the backbone entropy scale, corrected for the glycine backbone entropy loss taken from D'Aquino et al. (1996; Table 1). The correction does not influence the ranking of the designed sequences, because it only results in a constant offset. The following parameters were used in the calculations for optimization of the entropy weighting factor: distance-dependent electrostatic term with ε = 40R (R is the interatomic distance), penalty for polar hydrogen burial of 2 kcal/mole, and surface-based nonpolar exposure penalty and nonpolar burial benefit of 0.0232 kcal/mole/Å^2. The following amino acids were allowed at the designed positions: Ala, Val, Phe, Ile, Leu, Tyr, Trp, Met, and Ser.

Weighting factor for the electrostatic term

The following parameters were used: entropy term with weighting factor of 2.3, penalty for polar hydrogen burial of 2 kcal/mole, and surface-based nonpolar exposure penalty and nonpolar burial benefit of 0.048 kcal/mole/Å^2. The following amino acids were allowed at the designed positions: Ala, Val, Leu, Ile, Phe, Tyr, Trp, Asp, Asn, Glu, Gln, Lys, Ser, Thr, His, Hsp, Arg, and Met.

Weighting factor for the polar hydrogen burial penalty

The following parameters were used: entropy term with weighting factor of 2.3, dielectric constant of 10.3 R, and surface-based nonpolar exposure penalty and nonpolar burial benefit of 0.048 kcal/mole/Å^2. The amino acids allowed at the designed positions were the same as for the optimization of the electrostatics term.

Computational design

The crystal structure of hGH (Brookhaven Protein Data Bank code 3HHR) was used as the starting point. The program BIOGRAF (Molecular Simulations Inc.) was used to generate hydrogens on the structure and to minimize it (50 steps of conjugate gradient minimization with the Dreiding II force field; Mayo et.al. 1990). Residues were classified as core, surface, or boundary using the RESCLASS program (Dahiyat and Mayo 1997a). The parameters not specified in the Results section are described in other work (Dahiyat and Mayo 1996, 1997a). An expanded version (Dahiyat and Mayo 1996) of the backbone-dependent rotamer library of Dunbrack and Karplus (1993) was used in all the calculations.

Cloning and expression

A gene for hGH was synthesized from partially overlapping oligonucleotides (~100 bases) that were extended and PCR amplified. Codon usage was optimized for *Escherichia coli*, and several restriction sites were incorporated to ease future cloning. These partial genes were cloned into a vector and transformed into *E. coli* for sequencing. Several of these gene fragments were then cloned into adjacent positions in an expression vector (pET17 or pET21) to form the full-length gene for hGH and transformed into *E. coli* for expression. Protein was expressed in *E. coli* in insoluble inclusion bodies, and its identity was confirmed by immunoblot of SDS-PAGE using a commercial mAb against hGH (Santa Cruz Biotechnology).

Refolding

The protein inclusion bodies were dissolved and washed consecutively using wash buffer A (100 mM Tris at pH 8, 2% Triton, 4 M urea, 5 mM EDTA, 0.5 mM DTT) and wash buffer B (100 mM Tris at pH 8, 0.5 mM DTT), and the solvents were removed by centrifuging at 20,000*g* for 30 min. The pellet was resuspended with extraction buffer (50 mM glycine, 0.0156 M NaOH, 5 mM glutathione reduced, 8 M GdnHCl at pH 9.6). The supernatant was dialyzed for 12 to 16 h against folding buffer A (50 mM glycine, 0.0156 M NaOH, 10% sucrose, 1 mM EDTA, 1 mM glutathione reduced, 0.1 mM oxidized glutathione, 4 M urea at pH 9.6). The supernant was dialyzed for 6 to 8 h in buffer B (60 mM Tris, 10% sucrose, 1 mM EDTA, 0.1 mM reduced glutathione, 0.01 mM oxidized glutathione at pH 9.6).

Purification

A size exclusion column (10 mm × 300 mm loaded with Superdex prep 75 resin purchased from Pharmacia) was loaded with protein and eluted at a flow rate of 0.8 mL/min using the column buffer (100 mM Na_2SO_4, 50 mM Tris at pH 7.5). The peaks were monitored at dual wavelengths of 214 and 280 nm. Albumin, carbonic anhydrate, cytochrome C, and aprotinin were used to calibrate the molecular size of proteins versus elution time. The monomeric peak that elutes around the expected elution time for each protein was collected for biophysical characterization. The proteins were >98% pure as judged by reversed-phase high performance liquid chromatography on a C_4 column (3.9 mm × 150 mm), with a linear acetonitrile-water gradient containing 0.1% TFE. The identities of all proteins were confirmed by comparing the molecular mass measured by mass spectrometry with the corresponding molecular mass calculated using the protein sequences.

Spectroscopic characterization

Protein samples were 50 μM in 50 mM sodium phosphate (pH 5.5). Concentrations were determined using ultraviolet spectrophotometry. Protein structure was assessed by circular dichroism. Circular dichroism spectra were measured on an Aviv 202DS spectrometer equipped with a Peltier temperature control unit using a 1-mm path length cell. Thermal stability was assessed by monitoring the temperature dependence of the circular dichroism signal at 222 nm. The data were collected every 2.5°C, with an averaging time of 5 sec and an equilibration time of 3 min. The T_m of each protein was derived from the derivative curve of the ellipticity at 222 nm versus temperature. T_m values were reproducible to within 2°C for the same protein at the concentrations used.

Cell proliferation assay

Cell proliferation assays were performed using an interleukin 3–dependent murine proB cell line, BAF/B03, stably transfected with the full-length human growth hormone receptor (Behncken et al. 1997) according to the method of Rowlinson et al. (1995, 1996). Cells were maintained in RPMI-1640 medium with 5% fetal calf serum (FCS), 1 μg/mL gentamicin, and 50 units/mL interleukin 3. In preparation for the assay, exponentially growing cells were washed twice in PBS and resuspended in hGH-free and phenol red-free RPMI-1640 media with 5% FCS and 1 μg/mL gentamicin. Serial diluted hGH was then added to 96-well microtiter plates containing 2.5 × 10^4 cells/well. After 24 h of incubation at 37°C in 5% CO_2, cell proliferation was quantified using the MTT assay. In each assay, the wild type and all three of the designed variants of hGH were tested in triplicate on the same plate. The entire assay was repeated three times. EC_{50} values were determined using KaleidaGraph (Synergy Software) by nonlinear least-squares fit of sigmoidal parts of the averaged curves to a four parameter equation:

$$OD = OD_{max} - (OD_{max} - OD_{min})/(1 + ([hGH]/EC_{50})^n)$$

as performed by Young et al. (1997).

Acknowledgments

We thank Professor Michael J. Waters, University of Queensland, Australia, for conducting the cell proliferation assays and Drs. Marie Ary and John Desjarlais for editing the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby

marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Abagyan, R. and Totrov, M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235: 983–1002.

Atwell, S., Ultsch, M., De Vos, A.M., and Wells, J.A. 1997. Structural plasticity in a remodeled protein-protein interface. *Science* 278: 1125–1128.

Behncken, S.N., Rowlinson, S.W., Rowland, J.E., Conway-Campbell, B.L., Monks, T.A., and Waters, M.J. 1997. Aspartate 171 is the major primate-specific determinant of human growth hormone: Engineering porcine growth hormone to activate the human receptor. *J. Biol. Chem.* 272: 27077–27083.

Clackson, T., Ultsch, M.H., Wells, J.A., and de Vos, A.M. 1998. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* 277: 1111–1128.

Creamer, T.P. 2000. Side-chain conformational entropy in protein unfolded states. *Proteins* 40: 443–450.

———. 2001. Conformational entropy in protein folding: A guide to estimating conformational entropy via modeling and computation. *Methods Mol. Biol.* 168: 117–132.

Creamer, T.P. and Rose, G.D. 1992. Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc. Natl. Acad. Sci.* 89: 5937–5941.

Dahiyat, B.I. 1999. In silico design for protein stabilization. *Curr. Opin. Biotechnol.* 10: 387–390.

Dahiyat, B.I. and Mayo, S.L. 1996. Protein design automation. *Protein Sci.* 5: 895–903.

———. 1997a. De novo protein design: Fully automated sequence selection. *Science* 278: 82–87.

———. 1997b. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* 94: 10172–10177.

Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. 1997. Automated design of the surface positions of protein helices. *Protein Sci.* 6: 1333–1337.

D'Aquino, J.A., Gomez, J., Hilser, V.J., Lee, K.H., Amzel, L.M., and Freire, E. 1996. The magnitude of the backbone conformational entropy change in protein folding. *Proteins* 25: 143–156.

Desmet, J., Demaeyer, M., Hazes, B., and Lasters, I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356: 539–542.

de Vos, A.M., Ultsch, M., and Kossiakoff, A.A. 1992. Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science* 255: 306–312.

Dunbrack, Jr., R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* 230: 543–574.

Filikov, A.V. and James, T.L. 1998. Structure-based design of ligands for protein basic domains: Application to the HIV-1 Tat protein. *J. Comput Aided Mol. Des.* 12: 229–240.

Filikov, A.V., Mohan, V., Vickers, T.A., Griffey, R.H., Cook, P.D., Abagyan, R.A., and James, T.L. 2000. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput. Aided Mol. Des.* 14: 593–610.

Goldstein, R.F. 1994. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.* 66: 1335–1340.

Gordon, D.B., Marshall, S.A., and Mayo, S.L. 1999. Energy functions for protein design. *Curr. Opin. Struct. Biol.* 9: 509–513.

Hellinga, H.W. and Richards, F.M. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci.* 91: 5803–5807.

Hindmarsh, P.C. and Brook, C.G. 1987. Effect of growth hormone on short normal children. *Br. Med. J. (Clin. Res. Ed.)* 295: 573–577.

Jorgensen, K.D. 1987. Comparison of the pharmacological properties of pituitary and biosynthetic human growth hormone: Demonstration of antinatriuretic/antidiuretic and barbital sleep effects of human growth hormone in rats. *Acta Endocrinol. (Copenh)* 114: 124–131.

Kono, H., Nishiyama, M., Tanokura, M., and Doi, J. 1998. Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on side-chain packing. *Protein Eng.* 11: 47–52.

Kuhlman, B. and Baker, D. 2000. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* 97: 10383–10388.

Mayo, S.L., Olafson, B.D., and Goddard III, W.A. 1990. Dreiding: A generic force field for molecular simulations. *J. Phys. Chem.* 94: 8897–8909.

Moore, J.A., Pletcher, S.A., and Ross, M.J. 1986. Absorption enhancement of growth hormone from the gastrointestinal tract of rats. *Int. J. Pharm.* 34: 35–43.

Pearlman, R. and Bewley, TA. 1993. Stability and characterization of human growth hormone. *Pharm. Biotechnol.* 5: 1–58.

Pickett, S.D. and Sternberg, M.J. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* 231: 825–839.

Remmele, Jr., R.L., Nightlinger, N.S., Srinivasan, S., and Gombotz, W.R. 1998. Interleukin-1 receptor (IL-1R) liquid formulation development using differential scanning calorimetry. *Pharm. Res.* 15: 200–208.

Rowlinson, S.W., Barnard, R., Bastiras, S., Robins, A.J., Brinkworth, R., and Waters, M.J. 1995. A growth hormone agonist produced by targeted mutagenesis at binding site 1: Evidence that site 1 regulates bioactivity. *J. Biol. Chem.* 270: 16833–16839.

Rowlinson, S.W., Waters, M.J., Lewis, U.J., and Barnard, R. 1996. Human growth hormone fragments 1–43 and 44–191: In vitro somatogenic activity and receptor binding characteristics in human and nonprimate systems. *Endocrinology* 137: 90–95.

Schrier, J.A., Kenley, R.A., Williams, R., Corcoran, R.J., Kim, Y., Northey, Jr., R.P., D'Augusta, D., and Huberty, M. 1993. Degradation pathways for recombinant human macrophage colony-stimulating factor in aqueous solution. *Pharm. Res.* 10: 933–944.

Sternberg, M.J. and Chickos, J.S. 1994. Protein side-chain conformational entropy derived from fusion data: Comparison with other empirical scales. *Protein Eng.* 7: 149–155.

Stites, W.E. and Pranata, J. 1995. Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone. *Proteins* 22: 132–140.

Sundstrom, M., Lundqvist, T., Rodin, J., Giebel, L.B., Milligan, D., and Norstedt, G. 1996. Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 Å resolution. *J. Biol. Chem.* 271: 32197–32203.

Young, D.C, Zhan, H., Cheng, Q.L., Hou, J., and Matthews, D.J. 1997. Characterization of the receptor binding determinants of granulocyte colony stimulating factor. *Protein Sci.* 6: 1228–1236.

Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening

PEIZHI LUO, ROBERT J. HAYES, CHERYL CHAN, DIANE M. STARK,
MARIAN Y. HWANG, JONATHAN M. JACINTO, PADMAJA JUVVADI,
HELEN S. CHUNG, ANIRBAN KUNDU, MARIE L. ARY, AND BASSIL I. DAHIYAT
Xencor, Inc., Monrovia, California 91016, USA

(RECEIVED November 13, 2001; FINAL REVISION February 11, 2002; ACCEPTED February 13, 2002)

Abstract

Granulocyte-colony stimulating factor (G-CSF) is used worldwide to prevent neutropenia caused by high-dose chemotherapy. It has limited stability, strict formulation and storage requirements, and because of poor oral absorption must be administered by injection (typically daily). Thus, there is significant interest in developing analogs with improved pharmacological properties. We used our ultrahigh throughput computational screening method to improve the physicochemical characteristics of G-CSF. Improving these properties can make a molecule more robust, enhance its shelf life, or make it more amenable to alternate delivery systems and formulations. It can also affect clinically important features such as pharmacokinetics. Residues in the buried core were selected for optimization to minimize changes to the surface, thereby maintaining the active site and limiting the designed protein's potential for antigenicity. Using a structure that was homology modeled from bovine G-CSF, core designs of 25–34 residues were completed, corresponding to 10^{21}–10^{28} sequences screened. The optimal sequence from each design was selected for biophysical characterization and experimental testing; each had 10–14 mutations. The designed proteins showed enhanced thermal stabilities of up to 13°C, displayed five- to 10-fold improvements in shelf life, and were biologically active in cell proliferation assays and in a neutropenic mouse model. Pharmacokinetic studies in monkeys showed that subcutaneous injection of the designed analogs results in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment. These results show that our computational method can be used to develop improved pharmaceuticals and illustrate its utility as a powerful protein design tool.

Keywords: Protein design; computational screen; stability; cytokines; granulocyte-colony stimulating factor

Many techniques have been used in the design of new and improved proteins. In vitro directed evolution methods such as phage display, DNA shuffling, and error-prone PCR are widely used. Rational design approaches continue to be applied, and strategies that combine both are now being used. Successful designs include enzymes (Chen and Arnold 1991; Stemmer 1994; Zhao et al. 1998) and other proteins (Crameri et al. 1996), as well as therapeutically useful proteins such as hormones and cytokines (Lowman and Wells 1993; Heikoop et al. 1997; Grossmann et al. 1998; Chang et al. 1999). The experimental techniques involve the generation and screening of libraries of random protein sequences. However, the number of sequences that can be screened experimentally is limited (about 10^{14} for library panning and 10^7 for high throughput screening). Libraries of this size allow for the simultaneous modification of only about 10 residues.

Reprint requests to: Bassil I. Dahiyat, Xencor, Inc., 111 W. Lemon Avenue, Monrovia, California 91016, USA; e-mail: baz@xencor.com; fax: (626) 256-3562.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.4580102.

Computational methods have also been used that perform in silico screening of protein sequences (Hellinga and Richards 1994; Desjarlais and Handel 1995; Dahiyat and Mayo 1996, 1997a; Street and Mayo 1999; Jiang et al. 2000; Kraemer-Pecore et al. 2001; Pokala and Handel 2001). Exploiting the efficiency and speed of computers, these methods can randomly screen a vast number of sequences (up to 10^{80}), allowing for the simultaneous consideration and modification of more than 60 residues. Searching such large sequence spaces drastically improves the possibility of finding novel protein sequences with improved properties.

Investigators have recently developed a computational screening method that finds the optimal sequence for a defined three-dimensional structure, allowing all or part of the sequence to change (Dahiyat and Mayo 1996). This method, termed Protein Design Automation (PDA), scores the fit of sequences to the three-dimensional structure using physical-chemical potential functions that model the energetic interactions of protein atoms, including steric, solvation, and electrostatic interactions. PDA couples these potential functions with a highly efficient search algorithm to accurately screen up to 10^{80} sequences. Because the screening is performed in silico, multiple simultaneous mutations can be made, and novel sequences that are very different from wild type can be discovered. The method has been validated by numerous experimental tests and has resulted in the design of new proteins with improved stability and conformational specificity, and novel activity (Dahiyat and Mayo 1996, 1997a; Malakauskas and Mayo 1998; Strop and Mayo 1999; Shimaoka et al. 2000; Bolon and Mayo 2001; Marshall and Mayo 2001).

PDA also has the advantage of being able to control the location and type of mutations. For example, the design can be limited to the hydrophobic core. Mutations in the core can produce significant improvements in protein stability but do not change binding epitopes on the surface of the molecule. Thus, the molecular surface can be kept identical to the native structure, retaining biological activity and limiting toxicity and antigenicity. This feature is particularly important in the design of therapeutic proteins.

We wanted to take advantage of these features of PDA and explore its utility in the design of improved pharmaceuticals. We therefore used PDA as an ultrahigh throughput screen for improved analogs of a therapeutic protein, granulocyte-colony stimulating factor (G-CSF). G-CSF is a hematopoietic growth factor of 174 residues that induces differentiation and proliferation of granulocyte-committed progenitor cells. It is used clinically to treat cancer patients and alleviate the neutropenia induced by high-dose chemotherapy. G-CSF belongs to the class of long-chain four-helix bundle cytokines that bind asymmetrically to homodimeric complexes of cell-surface receptors to initiate an intracellular signaling cascade. Their structural similarity allows the design strategy chosen for G-CSF to be immediately applicable to the other four-helix bundle cytokines (human growth hormone, erythropoietin, the interleukins, and interferon-α/β—all clinically important compounds) and thus broadens the potential impact of the results.

Although the cytokines are functionally very efficacious, their pharmacological properties are not ideal. For example, G-CSF, like most proteins, is not absorbed orally to any significant extent and must be administered by frequent (daily) injections throughout the course of treatment. It also has limited stability and strict formulation and storage requirements, including the need to be kept refrigerated. Thus, there is significant interest in developing analogs with improved pharmacological properties.

We sought to use PDA to improve the physicochemical characteristics of G-CSF. Improving these properties can make a molecule more robust, enhance its shelf life, or make it more amenable to use in alternate delivery systems and formulations. It can also affect clinically important features such as pharmacokinetics and result in a drug that is safer for human use. Our design strategy was to optimize the core to improve the stability and solution properties of G-CSF while preserving receptor binding and biological activity.

The template structure used for in silico screening was a homology model of human G-CSF in which the human sequence was mapped onto bovine G-CSF. We designed several novel core sequences, cloned and expressed them, characterized their stabilities, tested them for functional activity both in vitro and in vivo, and studied their pharmacokinetics in monkeys. The designed proteins showed enhanced thermal stabilities, displayed five- to 10-fold improvements in shelf life, and were biologically active both in cell proliferation assays and in a neutropenic mouse model. Subcutaneous injection of the most stable variant in monkeys also resulted in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment. These results indicate that PDA has great potential as a powerful in silico tool in the design of improved pharmaceutical proteins.

Results and Discussion

Homology modeling

The crystal structure of bovine G-CSF (PDB record 1bgc) (Lovejoy et al. 1993) was used as the starting point for modeling because the crystal structure of human G-CSF (PDB record 1rhg) (Hill et al. 1993) is at a lower resolution and is missing key fragments, including a structurally important disulfide bond between positions 64 and 74. Bovine G-CSF is a good model for human G-CSF because the sequences are the same length and 142 of 174 amino acids are identical (82%). The residues that differ in the bovine sequence were replaced with the human residues for those positions, and the conformations of the replaced side chains were optimized using PDA. Most of the replaced residues were solvent exposed, thereby introducing little strain into the structure and allowing typical PDA parameters to be used for conformation optimization. One substitution, however, was at a buried site, G167V, and clashed sterically with a nearby disulfide bond. To accommodate the larger Val, the side-chain conformation at this position was optimized using a less restrictive van der Waals scale factor (0.6 instead of 0.9). The entire structure was then briefly minimized to relax the strain. The final structure that served as the template for all the designs is shown in Figure 1.

Core designs

Unlike many experimental sequence screening methods, PDA allows control over which residues are allowed to change. Core residues were selected because optimization of these positions can improve stability yet minimize changes to the molecular surface, thus limiting the designed protein's potential for antigenicity. Ala scanning studies of G-CSF indicate one or two binding sites on the protein surface that are probably responsible for granulopoietic activity (Reidhaar-Olson et al. 1996; Young et al. 1997) (Fig. 1). Although recent crystallographic studies of G-CSF complexed to its receptor show only one binding site in a novel 2:2 complex (Horan et al. 1996; Aritomi et al. 1999), both sites were avoided in the core designs to ensure preservation of function.

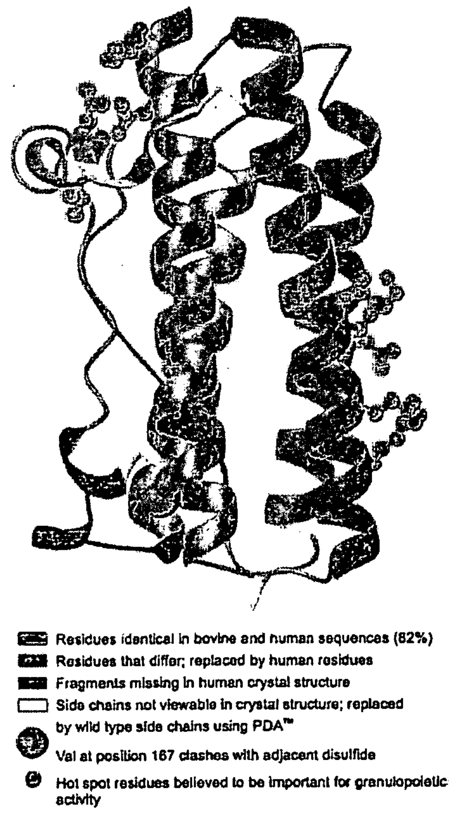

▨ Residues identical in bovine and human sequences (82%)
▨ Residues that differ; replaced by human residues
■ Fragments missing in human crystal structure
□ Side chains not viewable in crystal structure; replaced by wild type side chains using PDA™
● Val at position 167 clashes with adjacent disulfide
● Hot spot residues believed to be important for granulopoietic activity

Fig. 1. Template structure of hG-CSF used for Protein Design Automation (PDA) designs. The human sequence was homology modeled onto the bovine crystal structure (PDB record 1bgc). The residues that differ in the bovine sequence or were not present in the bovine crystal structure were replaced with the residues from the human sequence. The conformations of the replaced side chains were optimized using PDA (the larger Val at position 167 was optimized using a less restrictive van der Waals scale factor), and the entire structure was energy minimized for 50 steps.

Two PDA design calculations were run: a deep core design that included residues deeply buried in the interior of the protein and an expanded core design (exp_core) that also included less buried peripheral core residues. The deep core design had 26 core positions that were allowed to vary (shown yellow and gold in Fig. 2), whereas exp_core had 34 (shown yellow and turquoise in Fig. 2). Only hydrophobic amino acids were considered at the variable core positions. These included Ala, Val, Ile, Leu, Phe, Tyr, and Trp. Gly was also allowed for the variable positions that had Gly in the bovine wild-type structure (positions 28, 149, 150, and 167). Met and Pro were not allowed.

Optimal sequences

The optimal sequences selected by PDA are also shown in Figure 2. The optimal sequence from the deep core design had 10 mutations (named core10), and the optimal exp_core sequence had 11 (named exp_core11); thus, 33%–38% of the variable residues changed their identities. Eight of the mutated positions changed to the same amino acid in both designs. Changing the set of design positions can significantly impact the amino acid selected at a given position. For example, in the deep core design, Leu89 retains the same amino-acid identity and conformation as wild type. However, in the exp_core design, when Leu92 is also allowed to vary, both positions (Leu89 and Leu92) mutate to Phe, indicating a coupling between these two core residues. The modeled structure of the sequence selected in the deep core design (core10) is shown in Figure 3.

Native human G-CSF (met hG-CSF) and the optimal sequence from each of the core designs were cloned, expressed in *Escherichia coli*, and purified for experimental studies.

Thermal stability

The far-ultraviolet (UV) circular dichroism (CD) spectra for met hG-CSF and the designed proteins were nearly identical to each other and to published spectra for met hG-CSF (Reidhaar-Olson et al. 1996; Young et al. 1997), indicating highly similar secondary structure and tertiary folds (data

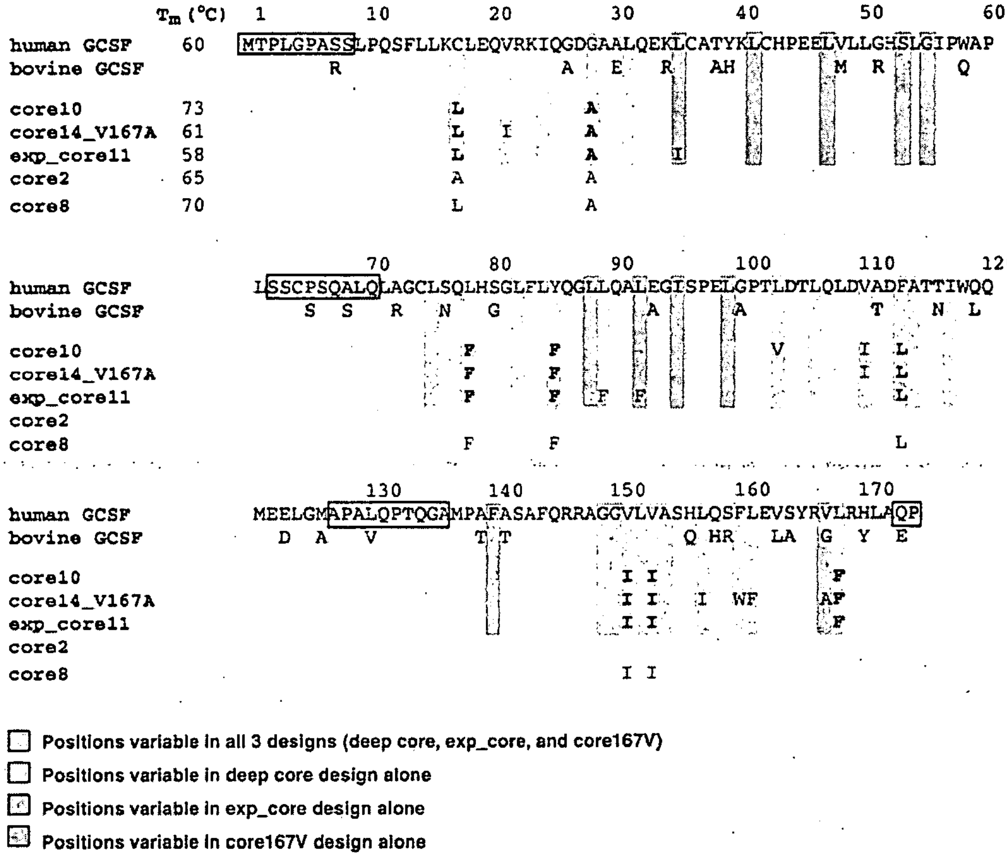

Tm (°C)
human GCSF 60
bovine GCSF
core10 73
core14_V167A 61
exp_core11 58
core2 65
core8 70
□ Positions variable in all 3 designs (deep core, exp_core, and core167V)
□ Positions variable in deep core design alone
□ Positions variable in exp_core design alone
□ Positions variable in core167V design alone

Fig. 2. Sequences of hG-CSF analogs. Native human and bovine sequences are shown at the *top*. The fragments missing in the crystal structure of the human sequence are shown boxed. Variable positions are colored. The deep core design had 26 variable positions, exp_core had 34, and core167V had 25. The optimal sequence from each design is shown. Letters indicate core residues that mutated relative to native hG-CSF; blanks indicate no change. Positions that changed to the same amino acid in all three core designs are indicated in bold. Core2 and core8 sequences were not obtained from PDA calculations but were derived by reverting some of the core10 mutations to wild type. Melting temperatures (T_ms) obtained for the designed proteins are also shown.

not shown). Thermal denaturation was monitored at 222 nm, and the melting temperatures (T_ms) were derived from the derivative curve of the ellipticity at 222 nm versus temperature (Fig. 4). Thermal denaturation of G-CSF and its variants is irreversible; however, T_m can be used to quickly assess the relative stability of different mutants. Stability under storage conditions, which is more relevant clinically, was evaluated with shelf-life studies (see below).

The T_m for met hG-CSF was 60°C, identical to that reported in other studies (Kolvenbach et al. 1997). Core10 showed an increase in stability of 13°C, whereas the T_m of exp_core11 was very similar to wild type (Fig. 2 and Fig. 4). The increased stability seen with core10 may be attributable to improved packing interactions and optimized hydrophobic burial of side-chains. Other possibilities include decreased aggregation resulting from elimination of the free cysteine at position 17. The Gly to Ala mutation at position 28 caused a significant improvement in helical propensity that could also be the source of the improved stability.

Identifying critical mutations using derived sequences

To differentiate between these possibilities, two additional sequences derived from the core10 mutant sequence were made and their T_ms measured. One of these (core8) was identical to core10 except that two mutations distant from the others were reverted to wild type (L103V and V110I). These were the two positions that did not mutate in exp_core11. The T_m of core8 was 70°C, similar to core10, indicating that the mutations at 103 and 110 were not responsible for core10's improved stability.

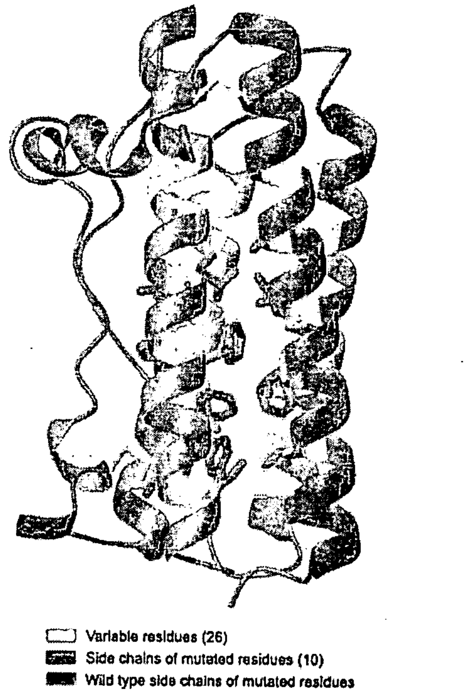

□ Variable residues (26)
▨ Side chains of mutated residues (10)
■ Wild type side chains of mutated residues

Fig. 3. Modeled structure of hG-CSF analog (core10) obtained from deep core design. Twenty-six core residues were allowed to vary; computational screening with PDA resulted in 10 mutations: C17L, G28A, L78F, Y85F, L103V, V110I, F113L, V151I, V153I, and L168F.

To determine the importance of the other mutations, another sequence was made (core2) that contained only two of the core10 mutations, G28A and C17A; all other residues were identical to wild type (Fig. 2). The T_m of core2 was 5°C higher than wild type, indicating that improvements in helical propensity and the elimination of a free cysteine are important for heightened thermostability. The remainder of the increase in T_m seen for core10 may be attributable to improved packing interactions and increased hydrophobic burial.

Storage stability

Increased shelf life is important for distribution and storage and is a desirable feature for G-CSF and other protein drugs. Because aggregation and chemical degradation are the predominant mechanisms of inactivation of G-CSF (Herman et al. 1996), shelf life was estimated by incubating the proteins at elevated temperature and then using size-exclusion chromatography to observe the disappearance of monomeric protein. Chemical degradation was estimated using reverse phase chromatography (data not shown). Core2 and core10 showed five- and 10-fold improvements in storage stability, respectively, at 50°C (Fig. 5). Rate constants were determined by a first order exponential fit of the fraction monomer remaining/time curves using KaleidaGraph (Synergy Software).

Biological activity

Granulopoietic activity was determined in vitro by quantitating cell proliferation as a function of protein concentration in murine lymphoid cells transfected with the gene for the human G-CSF receptor. The designed proteins were as

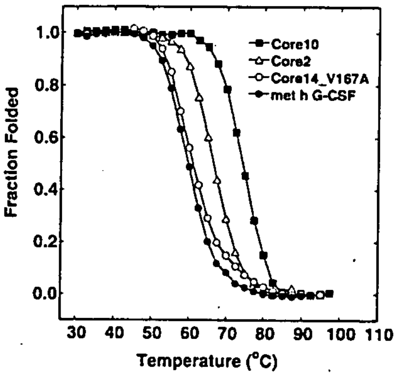


Fig. 4. Thermal stability of hG-CSF analogs. Thermal stability was assessed by monitoring the temperature dependence of the circular dichroism spectral signal at 222 nm. Melting temperatures (T_ms) were derived from the derivative curve of the ellipticity at 222 nm versus temperature. Core10 and core2 showed increases in T_m of 13°C and 5°C, respectively, over native met hG-CSF.

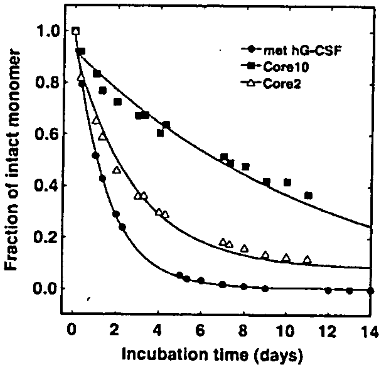


Fig. 5. Shelf life of hG-CSF analogs. Shelf life was estimated by incubating the proteins at elevated temperature (50°C) and using size exclusion chromatography to observe disappearance of monomeric protein. Rate constants were determined by a first order exponential fit of the fraction monomer remaining/time curves. Core2 and core10 showed five- and 10-fold improvements in storage stability, respectively, over met hG-CSF controls.

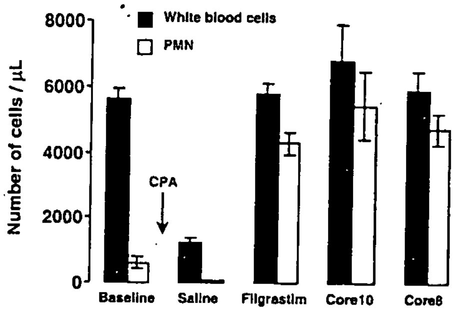


Fig. 6. In vivo granulopoietic activity of hG-CSF analogs. Mice were rendered neutropenic with a single intraperitoneal injection of 200 mg/kg cyclophosphamide (CPA). Beginning 24 h later and for 4 consecutive days, the mice were given a daily intravenous injection of 100 µg/kg of native hG-CSF (filgrastim, Amgen), an hG-CSF analog, or saline. On day 5, granulopoietic activity was determined by counting the number of white blood cells and polymorphonuclear neutrophils (PMN). The designed analogs (core8 and core10) were as effective as controls in eliciting a granulopoietic response.

active as wild-type hG-CSF (data not shown). The designed analogs were also as effective as wild type in increasing white blood cell and polymorphonuclear neutrophil levels in the neutropenic mouse (Fig. 6). Neutropenia, characterized by an abnormally low level of neutrophils in the blood, was induced by injection of cyclophosphamide. Reversal of this effect by the designed analogs shows that granulopoietic activity was also retained in vivo.

Pharmacokinetics

The pharmacokinetics of core10 and native hG-CSF (filgrastim, Amgen) was studied in cynomolgus monkeys after a single subcutaneous or intravenous injection of 5 µg/kg and after daily subcutaneous injections of 5 µg/kg for 28 d. Analysis of the serum concentration-time curves shows that subcutaneous injection of the designed analog results in greater systemic exposure (area under concentration-time curve, AUC) than the same dose of wild-type hG-CSF (Fig. 7B). This was true after a single dose on day 1 (78.8 vs. 54.6 ng-h/mL, data not shown), as well as on day 28 (37.2 vs. 17.4 ng-h/mL). There were no measurable differences in serum half-life. In the intravenous study, however, the half-life of core10 was three-fold shorter (1 vs. 3 h), and the AUC was significantly less (54.7 vs. 117.4 ng-h/mL), indicating that core10 is cleared faster (Fig. 7A). Taken together, these data indicate that the designed analog is absorbed more quickly from the subcutaneous compartment (absorption could not be measured directly given the small number of data points at early times). Improved absorption may be attributable to decreased aggregation or association of the designed protein. The increased monomer lifetime and decreased aggregation seen in our shelf-life studies and the improved thermal stability of the native conformation observed for core10 indicate a decrease in aggregation in the subcutaneous compartment. This possibility is supported by the fact that other protein therapeutics engineered for reduced aggregation also show faster absorption rates. For example, insulin Lispro and other rapid-acting insulin analogs that were designed to decrease their tendency to self-associate are absorbed faster than regular insulin after subcutaneous injection (Howey et al. 1994; Home et al. 1999).

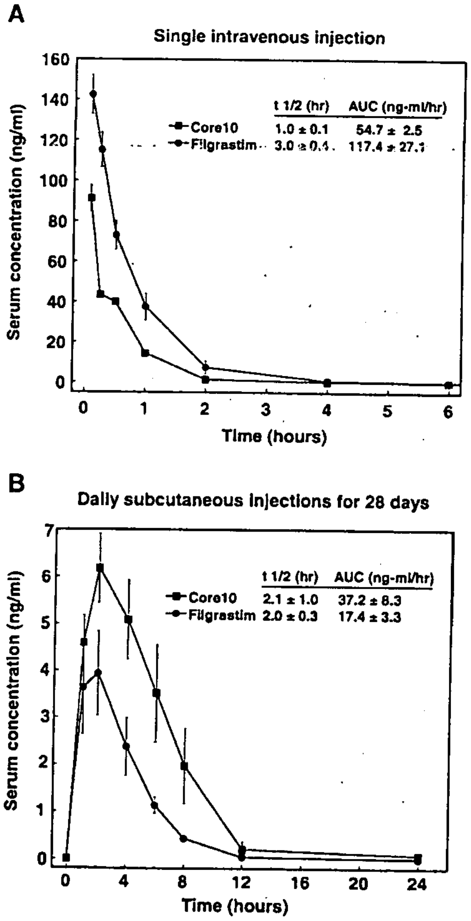


Fig. 7. Pharmacokinetics of hG-CSF analogs. Plasma concentrations of a designed hG-CSF analog or wild-type hG-CSF (filgrastim, Amgen) were determined after administration in cynomolgus monkeys. (*A*) Animals were given a single intravenous injection of 5 µg/kg or (*B*) daily subcutaneous injections of 5 µg/kg for 28 d. Noncompartmental analysis of the serum concentration-time curves shows that subcutaneous injections of the core10 analog resulted in greater systemic exposure (area under concentration-time curve, AUC) than the same dose of wild-type hG-CSF, whereas there was no change in serum half-life ($t_{1/2}$). In the intravenous study, the AUC was significantly less and the $t_{1/2}$ three-fold shorter, indicating that core10 was cleared faster.

Comparison to published G-CSF variants

In vitro and cassette mutagenesis studies have shown that alterations of the N-terminal region of G-CSF can lead to improved granulopoietic activity (Kuga et al. 1989; Okabe et al. 1990). Point mutations at Cys17 have also been found to affect shelf life; replacement with Ala led to an increase, Ser had no effect, and large residues (Ile, Tyr, Arg) led to a decrease (Ishikawa et al. 1992). In contrast, our core10 sequence, which has a large residue (Leu) at this position, showed an improved shelf life. This may be explained by the observation that in a Cys17Leu point mutant, Leu's side chain would clash with the aromatic ring of the nearby Phe at position 113. This steric clash does not occur in core10, however, because the Phe at 113 is replaced by Leu and, in compensation for this change, two nearby Leu's become Phe's (at positions 78 and 168). Thus, multiple mutations allow complementary repacking of the hydrophobic core in the core10 mutant and may be responsible for its enhanced stability and shelf life.

Significant improvements in thermal stability were also observed when the seven helical Gly residues in G-CSF were replaced with Ala to form point, double, and triple mutants (Bishop et al. 2001). Substitutions at positions 26, 28, 149, and 150 were the most effective. The investigators attributed the stabilizing effect to the enhancement in α-helical propensity associated with the Gly/Ala substitutions. These data support our suggestion that the heightened thermal stability seen with our mutants (which also contain a Gly/Ala substitution at position 28) is at least in part attributable to an improvement in helical propensity.

Probing the robustness of PDA with a homology modeled core position

As pointed out previously, the homology modeling of human G-CSF onto the bovine structure was straightforward for the most part because the replaced residues were primarily solvent exposed and no rearrangement of the backbone was necessary. The change at one core position, however, G167V, induced a steric clash and energy minimization of the entire protein was used to relieve the strain. We decided to assess the impact of this manipulation by doing an additional design (core167V) in which the variable residues were essentially the same as in the deep core design except that position 167 was also allowed to vary. We found that Val167 mutated to Ala (the other mutations were essentially the same as for core10). To probe the plasticity of the core, instead of using this PDA optimal sequence, which only had two mutations in this region, we ran experiments on another high-scoring sequence (core14_V167A) that had additional mutations (14 total, including L157I, F160W, and L161F). This sequence was chosen because it balanced an extensive number of mutations with a relatively high design score. Although it ranked 21st in the sequence energy list and was 2 kcal/mole less favorable than the optimal sequence, it was still biologically active and as stable as wild type (T_m of 61°C) (Figs. 2, 4). This indicates that optimization with PDA is fairly robust, and that the protein core can be quite plastic and can accommodate large changes without sacrificing stability or function.

Conclusions

PDA is a powerful ultrahigh throughput computational screening method. Its ability to screen up to 10^{80} sequences and allow multiple simultaneous mutations significantly increases the likelihood of finding new and improved proteins. In this study, PDA was used to develop improved analogs for a therapeutically important protein, hG-CSF. The novel proteins showed enhanced thermal stabilities and shelf life while retaining biological activity. Analysis of the mutants and results obtained with derived sequences indicates that the heightened stability is attributable to improvements in helical propensity and the elimination of a free cysteine; improved core packing and optimized hydrophobic burial of side chains may also be important. Pharmacokinetic studies indicate that subcutaneous injection of the most stable variant results in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment.

These results show that PDA can be successfully applied to proteins of therapeutic interest. They also illustrate the value of its precise control over the site and type of mutations, allowing for the rational design of desired properties such as improved stability and pharmacokinetics and the elimination of undesirable ones such as toxicity and antigenicity. These features are particularly important in the design of therapeutic proteins. PDA thus has great potential as a powerful in silico tool for therapeutic protein design.

Materials and methods

Template structure preparation

The template structure for the designed proteins was produced by homology modeling using the crystal structure of bovine G-CSF (Brookhaven Protein Data Bank code 1bgc) as the starting point. The program BIOGRAF (Molecular Simulations Inc., San Diego, CA) was used to generate explicit hydrogens on the structure, which was then minimized for 50 steps using the conjugate gradient method and the Dreiding II force field (Mayo et al. 1990). The residues that differ in the bovine sequence or were not present in the bovine crystal structure were replaced with the human residues for those positions. The conformations of the replaced side chains were optimized using PDA (Dahiyat and Mayo 1997a,b), and the entire structure was minimized again for 50 steps. This minimized structure was used as the template for all the designs.

Protein design

Analogs of hG-CSF were designed by simultaneously optimizing residues in the buried core of the protein using PDA. The computational details, residue classification, potential functions, and parameters used for van der Waals interactions, solvation, and hydrogen bonding are described in previous work (Dahiyat and Mayo 1996, 1997a). An expanded version of the backbone-dependent rotamer library of Dunbrack and Karplus (Dunbrack and Karplus 1993) was used in all the calculations. The global optimum sequence from each design was selected for characterization and experimental testing, except for core167V in which the 21st ranked sequence was used. Calculations were generally performed overnight using 16 processors of an SGI Origin 2000 with 32 R10000 processors running at 195 MHz. The length of the runs varied from 1 to several hours of CPU time.

Cloning and expression

A gene for met hG-CSF was synthesized from partially overlapping oligonucleotides (~100 bases) that were extended and PCR amplified. Codon usage was optimized for *E. coli* and several restriction sites were incorporated to ease future cloning. These partial genes were cloned into a vector and transformed into *E. coli* for sequencing. Several of these gene fragments were then cloned into adjacent positions in an expression vector (pET17 or pET21) to form the full-length gene for met hG-CSF (528 bases) and transformed into *E. coli* for expression. Protein was expressed in *E. coli* in insoluble inclusion bodies and its identity was confirmed by immunoblot of SDS-PAGE using a commercial mAb against hG-CSF.

Refolding, purification, and storage

The protein inclusion bodies were solubilized in detergent and refolded in the presence of $CuSO_4$ to promote formation of native disulfide bonds (Lu et al. 1992). A size-exclusion column (10 mm x 300 mm loaded with Superdex prep 75 resin purchased from Pharmacia) was loaded with protein and eluted at a flow rate of 0.8 mL/min using the column buffer (100 mM Na_2SO_4, 50 mM Tris, pH 7.5). The peaks were monitored at dual wavelengths of 214 nm and 280 nm. Albumin, carbonic anhydrate, cytochrome C, and aprotinin were used to calibrate the molecular size of proteins versus elution time. The monomeric peak that elutes around the expected elution time for each protein was collected and the buffer was exchanged into 10 mM NaOAc at pH 4 for biophysical characterization. For long-term storage, a buffer of 5% sorbitol, 0.004% Tween 80, and 10 mM NaOAc at pH 4 was used. A pH of 4 was chosen for these buffers to be consistent with the commercial formulation of hG-CSF (Amgen), which was used as a control. The proteins were >98% pure as judged by reversed phase high performance liquid chromatography (HPLC) on a C4 column (3.9 mm x 150 mm) with a linear acetonitrile-water gradient containing 0.1% TFE. The identities of all proteins were confirmed by comparing the molecular mass measured by mass spectrometry with corresponding molecular mass calculated using the protein sequences.

Spectroscopic characterization

Protein samples were 50 μM in 50 mM sodium phosphate at pH 5.5. Concentrations were determined using UV spectrophotometry. Protein structure was assessed by CD. CD spectra were measured on an Aviv 202DS spectrometer equipped with a Peltier temperature control unit using a 1-mm path length cell. Thermal stability was assessed by monitoring the temperature dependence of the CD signal at 222 nm (Kolvenbach et al. 1997). A buffer of 10 mM NaOAc was used at pH 4.0 and data were collected every 2.5°C with an averaging time of 5 sec and an equilibration time of 3 min. Thermal denaturation curves were smoothed using KaleidaGraph. The melting temperature (T_m) of each protein was derived from the derivative curve of the ellipticity at 222 nm versus temperature. The T_m values were reproducible to within 2°C for the same protein at the concentrations used.

Storage stability

The storage stability of the designed proteins was assessed by incubation at both 37°C and 50°C under solution conditions identical to that used in the commercial formulation of hG-CSF (filgrastim, Amgen). Because aggregation and chemical degradation are the predominant mechanisms of inactivation of G-CSF (Herman et al. 1996), accelerated degradation was followed by observing the disappearance of monomeric protein with both size-exclusion and reverse-phase chromatography. Rate constants for shelf-life estimation were determined by a first-order exponential fit of the fraction monomer remaining/time curves using KaleidaGraph (Synergy Software).

Cell proliferation assay

Granulopoietic activity was measured by quantifying cell proliferation as a function of protein concentration using Ba/F3 (murine lymphoid) cells stably transfected with the gene encoding the human Class 1 G-CSF receptor (Avalos et al. 1995). Cell proliferation was detected by 5-bromo-2′-deoxyuridine (BrdU) incorporation quantified by a BrdU-specific ELISA kit (Boehringer Mannheim).

In vivo biological activity

Granulopoietic activity was determined in the neutropenic mouse (Hattori et al. 1990). C57BL/6 mice were rendered neutropenic with a single intraperitoneal injection of 200 mg/kg cyclophosphamide (CPA). Beginning 24 h later and for 4 consecutive days, the mice were given a daily intravenous injection of 100 μg/kg of an hG-CSF analog, met hG-CSF produced in our laboratory, clinically available hG-CSF (filgrastim, Amgen), or saline. On day 5, 6 h after the final dose, the animals were killed, blood samples were collected, and granulopoietic activity was determined by counting the number of white blood cells and polymorphonuclear neutrophils.

Pharmacokinetics

Plasma concentrations of a designed hG-CSF analog or wild-type hG-CSF (filgrastim, Amgen) were determined following administration in cynomolgus monkeys. Animals were given a single intravenous injection of 5 μg/kg or daily subcutaneous injections of 5 μg/kg for 28 d. In the intravenous study, blood samples were collected at 0 (predose), 5, 15, and 30 min and 1, 2, 4, 6, 8, 12, and 24 h postdosing. In the subcutaneous studies, blood samples were collected at 0 (predose), 1, 2, 4, 6, 8, 12, and 24 h postdosing on day 1 and day 28. All samples were immediately placed on wet ice and centrifuged at 28°C. The resultant plasma was then frozen and stored (−70°C). Plasma concentrations were determined using an enzyme-linked immunosorbent assay (Quantikine human G-CSF ELISA, R&D Systems, Minneapolis, MN), performed per manufacturers instructions except that samples were diluted in PBS, 5% nonfat dry milk, and 0.05% Tween 20, and the incubation was extended to overnight at 4°C. Plasma concentrations of the designed hG-CSF analog and filgrastim were estimated from their corresponding standard curves. Pharmacokinetic parameters were calculated by noncompartmental analysis. The terminal slope (λz) was estimated by linear regression through the last time points of the log concentration versus time curves and used to calculate the terminal half-life ($t_{1/2}$). The area under the curve from time of dosing through the last time point (AUC_{0-z}) was calculated by the linear trapezoid method.

Acknowledgments

We thank Dr. Belinda Avalos (Ohio State University) for kindly supplying the Ba/F3 cell line transfected with the hG-CSF receptor. We also thank Dr. Steven Adams (American College of Laboratory Animal Medicine) and LAB Preclinical Research Institute Inc., (Quebec, Canada) for conducting the monkey studies.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Aritomi, M., Kunishima, N., Okamoto, T., Kuroki, R., Ota, Y., and Morikawa, K. 1999. Atomic structure of the GCSF-receptor complex showing a new cytokine-receptor recognition scheme. *Nature* **401**: 713–717.

Avalos, B.R., Hunter, M.G., Parker, J.M., Ceselski, S.K., Druker, B.J., Corey, S.J., and Mehta, V.B. 1995. Point mutations in the conserved box 1 region inactivate the human granulocyte colony-stimulating factor receptor for growth signal transduction and tyrosine phosphorylation of p75c-rel. *Blood* **85**: 3117–3126.

Bishop, B., Koay, D.C., Sartorelli, A.C., and Regan, L. 2001. Reengineering granulocyte colony-stimulating factor for enhanced stability. *J. Biol. Chem.* **276**: 33465–33470.

Bolon, D.N. and Mayo, S.L. 2001. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* **98**: 14274–14279.

Chang, C.C., Chen, T.T., Cox, B.W., Dawes, G.N., Stemmer, W.P., Punnonen, J., and Patten, P.A. 1999. Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* **17**: 793–797.

Chen, K.Q. and Arnold, F.H. 1991. Enzyme engineering for nonaqueous solvents: Random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology* **9**: 1073–1077.

Crameri, A., Whitehorn, E.A., Tate, E., and Stemmer, W.P. 1996. Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.* **14**: 315–319.

Dahiyat, B.I. and Mayo, S.L. 1996. Protein design automation. *Protein Sci.* **5**: 895–903.

———. 1997a. De novo protein design: Fully automated sequence selection. *Science* **278**: 82–87.

———. 1997b. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* **94**: 10172–10177.

Desjarlais, J.R. and Handel, T.M. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci.* **4**: 2006–2018.

Dunbrack, R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins—an application to side-chain prediction. *J. Mol. Biol.* **230**: 543–574.

Grossmann, M., Leitolf, H., Weintraub, B.D., and Szkudlinski, M.W. 1998. A rational design strategy for protein hormone superagonists. *Nat. Biotechnol.* **16**: 871–875.

Hattori, K., Shimizu, K., Takahashi, M., Tamura, M., Oheda, M., Ohsawa, N., and Ono, M. 1990. Quantitative in vivo assay of human granulocyte colony-stimulating factor using cyclophosphamide-induced neutropenic mice. *Blood* **75**: 1228–1233.

Heikoop, J.C., van den Boogaart, P., Mulders, J.W., and Grootenhuis, P.D. 1997. Structure-based design and protein engineering of intersubunit disulfide bonds in gonadotropins. *Nat. Biotechnol.* **15**: 658–662.

Hellinga, H.W. and Richards, F.M. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci.* **91**: 5803–5807.

Herman, A.C., Boone, T.C., and Lu, H.S. 1996. Characterization, formulation, and stability of Neupogen (Filgrastim), a recombinant human granulocyte-colony stimulating factor. *Pharm. Biotechnol.* **9**: 303–328.

Hill, C.P., Osslund, T.D., and Eisenberg, D. 1993. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc. Natl. Acad. Sci.* **90**: 5167–5171.

Home, P.D., Barriocanal, L., and Lindholm, A. 1999. Comparative pharmacokinetics and pharmacodynamics of the novel rapid-acting insulin analogue, insulin aspart, in healthy volunteers. *Eur. J. Clin. Pharmacol.* **55**: 199–203.

Horan, T., Wen, J., Narhi, L., Parker, V., Garcia, A., Arakawa, T., and Philo, J. 1996. Dimerization of the extracellular domain of granuloycte-colony stimulating factor receptor by ligand binding: A monovalent ligand induces 2:2 complexes. *Biochemistry* **35**: 4886–4896.

Howey, D.C., Bowsher, R.R., Brunelle, R.L., and Woodworth, J.R. 1994. [Lys(B28), Pro(B29)]-human insulin. A rapidly absorbed analogue of human insulin. *Diabetes* **43**: 396–402.

Ishikawa, M., Iijima, H., Satake-Ishikawa, R., Tsumura, H., Iwamatsu, A., Kadoya, T., Shimada, Y., Fukamachi, H., Kobayashi, K., Matsuki, S., et al. 1992. The substitution of cysteine 17 of recombinant human G-CSF with alanine greatly enhanced its stability. *Cell Struct. Funct.* **17**: 61–65.

Jiang, X., Farid, H., Pistor, E., and Farid, R.S. 2000. A new approach to the design of uniquely folded thermally stable proteins. *Protein Sci.* **9**: 403–416.

Kolvenbach, C.G., Narhi, L.O., Philo, J.S., Li, T., Zhang, M., and Arakawa, T. 1997. Granulocyte-colony stimulating factor maintains a thermally stable, compact, partially folded structure at pH2. *J. Pept. Res.* **50**: 310–318.

Kraemer-Pecore, C.M., Wollacott, A.M., and Desjarlais, J.R. 2001. Computational protein design. *Curr. Opin. Chem. Biol.* **5**: 690–695.

Kuga, T., Komatsu, Y., Yamasaki, M., Sekine, S., Miyaji, H., Nishi, T., Sato, M., Yokoo, Y., Asano, M., Okabe, M., et al. 1989. Mutagenesis of human granulocyte colony stimulating factor. *Biochem. Biophys. Res. Commun.* **159**: 103–111.

Lovejoy, B., Cascio, D., and Eisenberg, D. 1993. Crystal structure of canine and bovine granulocyte-colony stimulating factor (G-CSF). *J. Mol. Biol.* **234**: 640–653.

Lowman, H.B. and Wells, J.A. 1993. Affinity maturation of human growth hormone by monovalent phage display. *J. Mol. Biol.* **234**: 564–578.

Lu, H.S., Clogston, C.L., Narhi, L.O., Merewether, L.A., Pearl, W.R., and Boone, T.C. 1992. Folding and oxidation of recombinant human granulocyte colony stimulating factor produced in *Escherichia coli*. Characterization of the disulfide-reduced intermediates and cysteine—serine analogs. *J. Biol. Chem.* **267**: 8770–8777.

Malakauskas, S. and Mayo, S. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**: 470–475.

Marshall, S.A. and Mayo, S.L. 2001. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**: 619–631.

Mayo, S.L., Olafson, B.D., and Goddard III, W.A. 1990. Dreiding: A generic forcefield for molecular simulations. *J. Phys. Chem.* **94**: 8897–8909.

Okabe, M., Asano, M., Kuga, T., Komatsu, Y., Yamasaki, M., Yokoo, Y., Itoh, S., Morimoto, M., and Oka, T. 1990. In vitro and in vivo hematopoietic effect of mutant human granulocyte colony-stimulating factor. *Blood* **75**: 1788–1793.

Pokala, N. and Handel, T.M. 2001. Review: Protein design—where we were, where we are, where we're going. *J. Struct. Biol.* **134**: 269–281.

Reidhaar-Olson, J.F., De Souza-Hart, J.A., and Selick, H.E. 1996. Identification of residues critical to the activity of human granulocyte colony-stimulating factor. *Biochemistry* **35**: 9034–9041.

Shimaoka, M., Shifman, J.M., Jing, H., Takagi, J., Mayo, S.L., and Springer, T.A. 2000. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* **7**: 674–678.

Stemmer, W.P. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**: 389–391.

Street, A.G. and Mayo, S.L. 1999. Computational protein design. *Structure Fold. Des.* **7**: R105–109.

Strop, P. and Mayo, S.L. 1999. Rubredoxin variant folds without iron. *J. Am. Chem. Soc.* **121**: 2341–2345.

Young, D.C., Zhan, H., Cheng, Q.L., Hou, J., and Matthews, D.J. 1997. Characterization of the receptor binding determinants of granulocyte colony stimulating factor. *Protein Sci.* **6**: 1228–1236.

Zhao, H., Giver, L., Shao, Z., Affholter, J.A., and Arnold, F.H. 1998. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16**: 258–261.

Rational design and engineering of therapeutic proteins

Shannon A. Marshall, Greg A. Lazar, Arthur J. Chirino and John R. Desjarlais

An increasing number of engineered protein therapeutics are currently being developed, tested in clinical trials and marketed for use. Many of these proteins arose out of hit-and-miss efforts to discover specific mutations, fusion partners or chemical modifications that confer desired properties. Through these efforts, several useful strategies have emerged for rational optimization of therapeutic candidates. The controlled manipulation of the physical, chemical and biological properties of proteins enabled by structure-based simulation is now being used to refine established rational engineering approaches and to advance new strategies. These methods provide clear, hypothesis-driven routes to solve problems that plague many proteins and to create novel mechanisms of action. We anticipate that rational protein engineering will shape the field of protein therapeutics dramatically by improving existing products and enabling the development of novel therapeutic agents.

Shannon A. Marshall
Greg A. Lazar
Arthur J. Chirino
John R. Desjarlais*
Xencor
111 W. Lemon Ave.
Monrovia
CA 91016, USA
tel: +1 626 305 5900
fax: +1 626 305 0350
*e-mail: jrd@xencor.com

▼ The exquisite specificity of biological therapeutics for their clinical targets has led to their continued development and application as medicines, despite competition from small molecule drugs. Several engineered protein therapeutics are currently being marketed (Table 1), and the annual sales of protein therapeutics are projected to exceed US$59 billion in 2010, which is twice the revenue generated in 2001 (http://www.pharmafile.com/Pharmafocus/Features/feature.asp?fID=281) [1]. For well-validated targets, naturally occurring protein interaction partners constitute preselected 'lead' compounds with high affinity and specificity. However, because natural proteins are not evolved for utilization as drugs, lead optimization is frequently beneficial for development of a protein therapeutic. Modifications can influence the mechanism of action, side effects and efficacy, and satisfy practical constraints such as production costs, intellectual property and dosing frequency.

A variety of strategies have emerged for modulating protein properties, such as efficacy, stability, specificity, immunogenicity and pharmacokinetics (PK). Mechanisms for altering these properties include manipulation of primary structure, incorporation of chemical and post-translation modifications and utilization of fusion partners. The most common route to optimization is site-directed mutagenesis, which is often performed in a brute force or trial-and-error manner. A smaller number of examples exist whereby semirational application of diversity methods, such as phage display, has been used to optimize a therapeutic candidate. Important recent developments are the creation and successful application of rational protein design methods and the determination of an increasing number of high-resolution protein structures.

For the purposes of this review, we define rational protein engineering as the hypothesis-driven manipulation of protein sequence and/or composition. The controlled modification of specific biophysical properties of proteins can potentially impact a variety of therapeutic features (Table 2). An important subset of rational engineering methods consists of approaches that utilize high-resolution, 3D structure information. The most sophisticated of these methods offers an extraordinary level of control over protein sequence and structure, a mechanism to explore sequence combinations that extends far beyond natural diversity, and the ability to couple multiple constraints algorithmically for

1359-6446/03/$ - see front matter ©2003 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(03)02610-2

Table 1. Engineered protein therapeutics on the market[a]

Name	Family	Company	Indication	Modification	Property
Proleukin® (aldesleukin)	IL-2	Chiron	Cancer	Mutated free cysteine	Decreased aggregation; improved bioavailability
Betaseron® (interferon beta-1b)	IFN-β	Berlex/Chiron	Multiple sclerosis	Mutated free cysteine	Decreased aggregation
Humalog® (insulin lispro)	Insulin	Eli Lilly	Diabetes	Monomer not hexamer	Fast acting
NovoLog® (insulin aspart)	Insulin	Novo Nordisk	Diabetes	Monomer not hexamer	Fast acting
Lantus® (insulin glargine)	Insulin	Aventis	Diabetes	Precipitates in dermis	Sustained release
Enbrel® (etanercept)	TNF receptor	Immunex/ Amgen/Wyeth	Rheumatoid arthritis	Fc fusion	Longer serum half-life; increased avidity
Ontak® (denileukin diftitox)	Diptheria toxin-IL-2	Seragen/Ligand	Cancer	Fusion	Targets cancer cells
PEG-Intron® (peginterferon alfa-2b)	IFN-α	Schering-Plough	Hepatitis	PEGylation	Increased serum half-life; weaker receptor binding
PEGasys® (peginterferon alfa-2a)	IFN-α	Roche	Hepatitis	PEGylation	Increased serum half-life; weaker receptor binding
Neulasta™ (pegfilgrastim)	G-CSF	Amgen	Leukopenia	PEGylation	Increased serum half-life
Oncaspar® (pegaspargase)	Asparaginase	Enzon	Cancer	PEGylation	Decreased immunogenicity; increased serum half-life
Aranesp® (darbepoetin alfa)	Epo	Amgen	Anemia	Additional glycosylation sites	Increased serum half-life; weaker receptor binding
Somavert® (pegvisomant)	Growth hormone	Genentech/ Seragen/ Pharmacia	Acromegaly	PEGylation; binding site mutations	Novel mode of action; increased serum half-life

Chiron (http://www.chiron.com); Berlex (http://berlex.com); Eli Lilly (http://www.lilly.com); Novo Nordisk (http://www.novonordisk.com); Aventis (http://www.aventis.com); Immunex/Amgen (http://www.amgen.com); Wyeth (http://www.wyeth.com); Seragen/Ligand (http://www.ligand.com); Schering-Plough (http://www.sch-plough.com); Roche (http://www.roche.com); Enzon (http://www.enzon.com); Genentech (http://www.genentech.com); Pharmacia (http://www.pharmacia.com).

[a] Abbreviations: G-CSF, granulocyte-colony stimulating factor; IFN-α, interferon α; IL-2, interleukin 2; PEG, polyethylene glycol; TNF, tumor necrosis factor.

simultaneous optimization of several protein properties. Furthermore, proven hypotheses can be reapplied to additional protein systems, thus saving discovery cost and time. Rational methods can be distinguished from those that rely on random sequence perturbations or combinations, such as the class of optimization techniques referred to as directed evolution, although some implementations of these methods have a rational component [2,3].

Physicochemical properties

The physical and chemical properties of protein therapeutics significantly determine their performance during development, manufacturing and clinical use. Many therapeutically interesting proteins are naturally expressed at low concentrations and are degraded rapidly. By contrast, fully developed protein therapeutics require high levels of solubility as well as retention of activity through purification, formulation, storage and administration. Several rational design and engineering strategies, such as those highlighted in Figue 1, have been developed to improve properties such as solubility and stability while maintaining desired biological activity.

Stability

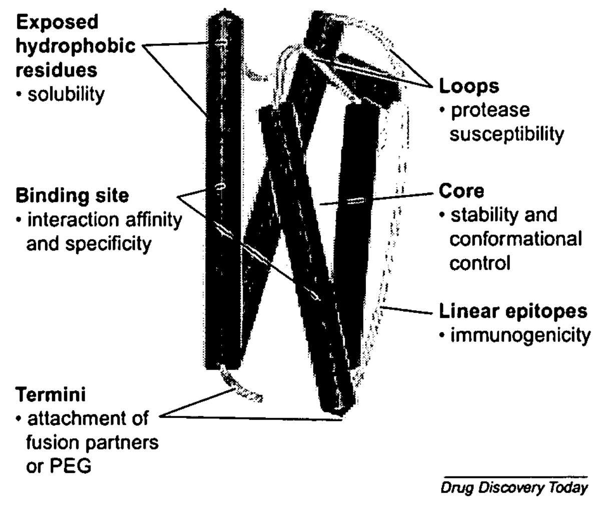
Protein therapeutics are exposed to a variety of stresses that can cause protein unfolding or degradation. Using rational optimization methods, proteins can be re-engineered

Table 2. The biophysical pr perties of proteins that can be optimized to obtain desired therapeutic outcomes[a]

	Enable discovery	Mechanism of action	Pharmaco-kinetics	Immuno-genicity	Route of administration	Cost of goods	Shelf life	Intellectual property
Stability	x				x	x	x	
Solubility	x		x	x	x	x	x	
Receptor binding affinity and specificity		x	x					
MHC binding affinity			x	x				
Oligomerization state		x	x		x			
Chemical modifications		x	x	x	x		x	x
Posttranslational modifications		x	x	x			x	x
Sequence diversity								x
Conformational state		x						

[a] Abbreviation: MHC, major histocompatibility complex.

such that their structure and activity are substantially more robust with respect to protease exposure, oxidative stress and changes in temperature, pH and solution conditions. One simple stabilization strategy is to replace free cysteines, thereby preventing the formation of unwanted intermolecular or intramolecular disulfide bonds. Cysteine to serine mutations have been introduced successfully into several therapeutic proteins, including granulocyte colony-stimulating factor (G-CSF) and interferon (IFN) βlb, resulting in a longer shelf life [4,5]. Cysteine to serine mutations have also been shown to increase the half-life of human fibroblast growth factor (FGF) [6]. Interestingly, each of the three FGF mutations decreases the thermal stability of the protein, probably because the introduced serines are substantially desolvated and are not positioned to form intramolecular hydrogen bonds. Rational approaches can identify the amino acids that are more precisely compatible with the local structural environment.


Exposed hydrophobic residues
• solubility
Binding site
• interaction affinity and specificity
Termini
• attachment of fusion partners or PEG
Loops
• protease susceptibility
Core
• stability and conformational control
Linear epitopes
• immunogenicity

Drug Discovery Today

Figure 1. Examples of rational design and engineering strategies. Many design strategies target specific residues or regions of a protein structure for optimization. However, it is important to note that modifications in any region of the protein could potentially affect a wide range of protein properties, emphasizing the importance of rational design methods that can simultaneously consider and optimize multiple parameters. Abbreviation: PEG, polyethylene glycol.

Dramatic improvements in the global stability of a protein can be obtained by optimizing intramolecular interactions. Early examples used rational computational design methods to optimize packing interactions and hydrophobic burial in the protein core [7–9]. Optimizing secondary structure propensity, hydrogen bonds and electrostatic interactions can also improve protein stability substantially [10–12]. More recently, these principles have been applied to the clinically relevant proteins G-CSF and human growth hormone (hGH) by using Protein Design Automation® (PDA™) technology (Box 1). The designed hGH variants are active in cell proliferation studies and are up to 16°C more thermostable than the wild type protein [13]. Optimized G-CSF variants with 10–14 mutations display enhanced thermal stability and five to tenfold increases in shelf life while maintaining the desired biological activity [14].

An additional stabilization strategy is reduction of proteolytic susceptibility. If a specific site in the protein is known to be especially prone to proteolysis, it can be modified so that it no longer matches the substrate specificity of the putative protease. The protease cleavage sites are often

Box 1. Xencor's PDA™ technology: a state of the art rational engineering platform

PDA™ technology, originally developed at Caltech [10,66,67] and further optimized at Xencor [13,14], couples computational design algorithms that generate quality sequence diversity with experimental high-throughput screening to discover proteins with improved properties. The computational component uses atomic level scoring functions, side chain rotamer sampling and advanced optimization methods to capture the relationships between protein sequence, structure and function accurately. Calculations begin with the 3D structure of the protein and a strategy to optimize one or more properties of the protein. PDA™ technology then explores the sequence space comprising all pertinent amino acids (including unnatural amino acids, if desired) at the positions targeted for design. This is accomplished by sampling conformational states of allowed amino acids and scoring them using a parameterized and experimentally validated function that describes the physical and chemical forces governing protein structure. Powerful combinatorial search algorithms are then used to search through the initial sequence space, which can constitute 10^{50} sequences or more, and quickly return a tractable number of sequences that are predicted to satisfy the design criteria. Useful modes of the technology span from combinatorial sequence design to prioritized selection of optimal single site substitutions. PDA™ technology has been applied to numerous systems including important pharmaceutical and industrial proteins and has a demonstrated record of success in protein optimization.

located in flexible loops; therefore, another approach is to introduce mutations that decrease flexibility. Thrombolytics are a class of protein therapeutics for which proteolytic susceptibility is especially important because many clotting factors are both activated and inactivated by specific proteases. For example, an engineered variant of coagulation factor VIIIa with increased resistance to proteolytic inactivation was generated by mutating two arginines required for cleavage by thrombin, factor Xa and activated protein C [15].

Solubility

Protein therapeutics are typically expressed, formulated and administered at high concentrations. Under such conditions many proteins form inclusion bodies during expression or aggregates after formulation. Improving the solubility of a protein can facilitate discovery efforts, whereas enabling soluble prokaryotic expression can reduce production costs dramatically and increase yields. It is far more critical to ensure the solubility of a protein therapeutic once it is administered. Aggregation can cause decreased activity, decreased bioavailability and increased immunogenicity. Several strategies have been applied successfully to reduce protein aggregation and enable soluble expression. Replacement of unpaired cysteine residues can prevent the formation of unwanted intermolecular disulfide bonds, as described above. Post-translational and chemical modifications, which are discussed in a later section, can also help to prevent aggregation. Substituting exposed nonpolar residues with polar residues can enable soluble expression and improve the solubility of the purified protein. This strategy was applied successfully to the A1 domain of cholera toxin, a powerful adjuvant. Of the six variants produced, one retained full biological activity and stability and also displayed a significant improvement in solubility [16]. Altering the net charge and isoelectric point (pI) of a protein can also affect its solubility. For example, a single chain antibody targeting renal cell carcinoma was altered to increase solubility by adding five glutamic acid residues to the C-terminus, thus lowering the pI from 7.5 to 6.1 [17]. Although there are a few examples of rational solubility engineering, the majority of the published successes in solubility optimization have been anecdotal. Until now solubility obstacles have been more or less considered to be formulation problems that can be surmounted with an exhaustive protein chemistry effort. We anticipate that systematic structure-guided optimization efforts will lead to the emergence of well-defined strategies that consistently yield proteins with improved solubility and minimally perturbed structure and function.

Pharmacokinetics

The first generation of protein therapeutics frequently suffered from poor PK. As our understanding of protein clearance processes improves, it becomes possible to rationally modify proteins to tailor their PK profiles. Properly controlling the serum concentration of a therapeutic protein over time can lead to improved efficacy and decreased side effects. In fact, improvements in PK properties can be so vital to the efficacy of a protein drug that they are often made at the expense of specific activity. In addition to eliminating proteolytic susceptibility (see above), several strategies have been developed to alter PK, including polyethlylene glycol (PEG) attachment (PEGylation), glycosylation, fusion to proteins with long serum half-lives, alteration of oligomerization state and modulation of receptor-mediated uptake and turnover. Knowledge of the dominant route or routes of elimination for a given protein can help significantly in determining which of these strategies will be the most appropriate. For low molecular weight protein therapeutics, kidney filtration dominates, encouraging modifications that increase the effective size. In the case of ligand-receptor systems PK often depends on the relative influence of receptor-mediated clearance versus renal clearance. Affinity and specificity modifications are a central component of many therapeutic optimization strategies, and thus receptor-mediated clearance might play an important role in the efficacy of many proteins, even when it is not considered explicitly.

Fusion proteins

In a straightforward application of molecular size manipulation, proteins covalently fused to themselves often display significantly improved PK profiles [18]. The PK of a therapeutic protein can be increased more dramatically through fusion to a protein that is known to have a long serum half-life, typically albumin or the Fc region of antibodies. Amgen and Wyeth's Enbrel® (etanercept), which is currently marketed for the treatment of rheumatoid arthritis, is a fusion protein consisting of the extracellular domain of p75 tumor necrosis factor receptor (TNFR) and the Fc domain of human IgG. Fc increases the serum half-life of Enbrel®, presumably by both increasing its size and mediating endosomal recycling (see below). Furthermore, because Fc is a dimer, Enbrel®'s affinity for TNF-α is 50- to 1000-fold higher than the affinity of monomeric TNFR [19]. Albumin fusions have been used to generate variants of the anticoagulant proteins hirudin [20] and barbourin [21]. An interesting twist on this approach is to tag proteins with a peptide sequence that specifically binds albumin. Addition of an albumin-binding peptide tag to the antitissue factor D3H44 Fab increases its half-life by approximately 40-fold [22].

Alteration of oligomerization state

The rate of absorption after injection can be affected by the molecular weight and solubility of a protein. An interesting example is provided by comparing wild type insulin, fast-acting insulin variants and sustained-release insulin variants. Native insulin forms a mixture of dimers and hexamers. The fast-acting insulin variants produced by Eli Lilly and Novo Nordisk, Humalog® (insulin lispro) and NovaLog® (insulin aspart) respectively, contain mutations that decrease oligomerization and, therefore, increase the rate of absorption. As a result, patients can administer these fast-acting insulin variants at mealtimes rather than 1 hour before as was required with native insulin. Long-acting insulin variants are used to maintain steady basal insulin levels. For example, the Aventis product Lantus® (insulin glargine) was engineered by increasing the pI to promote precipitation upon subcutaneous injection, thus slowing the rate of absorption [23].

PEGylation

PEG is a highly flexible and soluble polymer that has gained widespread scientific and regulatory acceptance as a chemical modification for therapeutic proteins. PEGylation improves PK predominantly by increasing the effective size of a protein, with most significant effects for proteins smaller than 70 kDa [24,25]. PEGylation can also reduce immunogenicity and aggregation [26]. Although a variety of chemistries exist [27,28] for coupling PEGs of various sizes to proteins, the greatest attachment specificity generally arises from PEGylation at the N-terminus or unpaired cysteines.

Several PEGylated protein therapeutics, such as Schering-Plough's PEG-Intron® (peginterferon alfa-2b) and Roche's PEGasys® (peginterferon alfa-2a), are currently on the market or in late-stage clinical trials. PEGasys® exhibits a 50- to 70-fold increase in serum half-life and substantially reduced variability in serum concentration [29]. A common negative effect of PEGylation, exemplified by both PEGylated IFNs [29,30], is a loss of specific activity. Future studies on these and other proteins should, therefore, focus on minimizing activity loss by optimizing the sites and sizes of PEG attachment rationally.

Glycosylation

Site-specific incorporation of glycosylation sites serves as an additional approach for improving PK. A notable example is Amgen's hyperglycosylated erythropoietin (Epo) variant Aranesp® (darbepoetin alfa), engineered to contain two additional N-linked glycosylation sites. The additional glycosylation increases the serum half-life threefold while reducing *in vitro* binding roughly fourfold [31]. Thus, Aranesp® is another example of how modification can improve *in vivo* efficacy, despite reducing specific activity. Accordingly, future efforts could benefit from using rational methods to identify N-linked or O-linked glycosylation sites that best maintain the structural and functional properties of the protein.

Endocytic trafficking

The PK of many proteins that bind cell-surface receptors can also be affected by endocytic trafficking. Cell-surface receptors and bound ligands are continually internalized by endocytosis. The receptors and ligands can be recycled back to the surface, degraded in lysozomes or transported across cells (e.g. from the apical membrane to the basolateral membrane). The fate of the ligand is often determined by the extent of association with the receptor within the endosome, although the relationship between processing and association is highly system dependent. The pH in endosomal compartments is lower than that of serum. Because protein–protein interactions are typically pH-dependent, many ligands are freed from their receptors as they proceed through the endosomal pathway. In some protein families released ligand is recycled, whereas ligand that remains bound is targeted for degradation. For example, the more tightly epidermal growth factor (EGF) family ligands bind EGF receptor at pH 6, then the lower the fraction of ligand that is recycled [32]. Recently Lauffenberger and colleagues have taken advantage of the pH dependence of endosome-mediated G-CSF turnover to rationally engineer variants with improved PK. Several residues involved in receptor binding were mutated to histidine, which is neutrally charged in serum but has a net positive charge in the acidic environment of endocytic vesicles. The variants were predicted to bind receptor normally at the cell surface but to release more effectively than the wild type after endocytosis. Two mutations were shown to have increased half-life and potency compared with wild type G-CSF [33]. This elegant example illustrates the ability of rational engineering methods to use accumulated biological knowledge to generate improved therapeutics. Another important example, discussed below in more detail, is the pH-dependent recycling of immunoglobulin Fc domains. In this case the effect is opposite: the pH drop purges antigen from the variable region while enhancing Fc binding to its receptor, thus enabling the antibody and its receptor to be recycled to the serum.

Affinity, specificity and conformational control

Rational design can be used to modify the affinity and specificity of interactions between a therapeutic protein and other biomolecules. In some cases, increasing the binding affinity for a target protein can produce an increase in biological activity. In other cases, it is possible to reduce undesired biological activities by decreasing the affinity for nontarget molecules. An example of affinity enhancement is the generation of superagonist variants of human thyrotropin (hTSH) by altering the net charge of the protein. The hTSH receptor has a net negative charge, and mutations that introduce positively charged residues or replace negatively charged residues in the peripheral loops of hTSH increase activity. The best variants show a 50,000-fold increase in receptor binding affinity and 1000-fold increase in *in vivo* activity [34,35].

The power of rational design is most impressive when it is used to generate novel mechanisms of action. For example, 4-helix bundle cytokines, including vascular endothelial growth factor (VEGF), hGH and interleukin-6 (IL-6), have been engineered to function as receptor antagonists rather than agonists. Most members of the 4-helix bundle cytokine family must form multiple protein–protein interactions at the cell surface to trigger signaling. VEGF, for example, forms homodimers that bind to two VEGF receptors, whereas IL-6 binds to a low-affinity IL-6 co-receptor and gp130. Antagonistic VEGF variants were designed as heterodimers, which contain one functional binding site per dimer [36]. An IL-6 superantagonist was generated by selecting mutations that disrupt binding to gp130 and incorporating mutations that result in increased affinity for the IL-6 co-receptor [37]. An especially interesting example of a designed cytokine antagonist is Genentech/Pharmacia's Somavert® (pegvisomant), a hGH variant that has recently successfully completed clinical trials for treatment of acromegaly. hGH contains two distinct receptor binding sites and dimerizes its receptor upon binding. Somavert® contains a point mutation at the second receptor binding site that blocks receptor dimerization [38] and eight additional mutations, identified by phage display, that increase the receptor-binding affinity of the first site [39].

Many proteins undergo conformational changes that are central to their function. In such cases, rational design methods can drive conformational equilibria towards the therapeutically desirable state. A notable example is the design of constitutively active and inactive integrin I domain variants. Integrin I domains can populate two dominant conformations: an 'open' conformation, which can bind intracellular adhesion molecule-1 (ICAM-1), and a 'closed' conformation, which has very low affinity for ICAM-1. The native protein rests in the closed conformation and converts to the open conformation during signaling. Springer and coworkers used two distinct strategies to generate conformationally locked integrin I domain variants. One approach introduced pairs of cysteines that form disulfide bonds compatible with either the closed or open conformation [40,41]. In the second approach, mutations were designed in the core of the domain that were computationally selected to stabilize the open conformation and disallow the closed state [42].

Immunogenicity

The potential for protein therapeutics to produce harmful immune responses is a significant barrier to the development and acceptance of protein drugs. The immune response is typically most severe for nonhuman proteins. For example, antibodies against streptokinase, a bacterially derived

Table 3. Engineered antibodies on the market[a]

Name	Company	Target	Indication	Type
Orthoclone OKT3® (muromonab-CD3)	Ortho Biotech/ Johnson & Johnson	CD3	Transplant rejection	Murine
ReoPro® (abciximab)	Centocor/Lilly	GPIIb/IIIa	Restenosis	Chimeric
Rituxan® (rituximab)	IDEC/Genentech	CD20	B-cell non-Hodgkins lymphoma	Chimeric
Simulect® (basiliximab)	Novartis	IL-2R	Transplant rejection	Chimeric
Remicade® (infliximab)	Centocor	TNF-α	Crohn's disease, rheumatoid arthritis	Chimeric
Zevalin® (ibritumomab tiuxetan)	IDEC/Schering AG	CD20	B-cell non-Hodgkins lymphoma	Chimeric
Zenapax® (daclizumab)	PDL/Roche	IL-2R	Transplant rejection	Humanized
Synagis® (palivizumab)	MedImmune	RSV F protein	Respiratory syncitial virus	Humanized
Herceptin® (trastuzumab)	Genentech	HER2/neu	Breast cancer	Humanized
Mylotarg® (gemtuzumab ozogamicin)	Celltech/Wyeth	CD33	Acute myeloid leukemia	Humanized
Campath® (alemtuzumab)	Millenium/ILEX	CD52	B-cell chronic lymphocytic leukemia	Humanized

Ortho Biotech (http://www.orthobiotech.com); Johnson & Johnson (http://www.jnj.com); Centocor (http://www.centocor.com); Eli Lilly (http://www.lilly.com); IDEC (http://www.idec.com); Genentech (http://www.genentech.com); Novartis (http://www.novartis.com); Schering Ag (http://www.schering.de/eng/); PDL (http://www.pdl.com); Roche (http://www.roche.com); MedImmune (http://www.medimmune.com); Celltech (http://www.celltechgroup.com); Wyeth (http://www.wyeth.com); Millenium (http://www.mlnm.com); ILEX (http://www.ilexonc.com).

[a] Abbreviations: GPIIb/IIIa, platelet glycoprotein IIb/IIIa; IL-2R , interleukin 2 receptor; RSV F, respiratory syncitial virus F; TNF-α, tumor necrosis factor α.

antithrombolytic used to treat myocardial infarction, not only neutralize the protein and reduce its efficacy, but also can elicit severe allergic reactions that effectively limit streptokinase therapy to one-time use. Yet even therapeutics based on human proteins can cause immune responses depending on the mode of administration (including dosage, frequency and route) and the solubility and stability of the formulated protein. Neutralizing antibodies have been observed against a variety of human proteins including insulin, factor VIII, IFNs, Epo and megakaryocyte growth and differentiation factor (MGDF). In some cases, for example with the multiple sclerosis drug IFN-β, efficacy is severely hindered due to neutralizing antibodies [43]. Devastating problems can result when elicited antibodies crossreact with endogenous protein. For example, clinical trials of MGDF were halted when crossreactive neutralizing antibodies to endogenous thrombopoietin caused reduced platelet counts (thrombocytopenia) in a small number of otherwise healthy volunteers. As another example, Johnson and Johnson's European formulation of Epo, Eprex® (epoetin alfa) has caused pure red blood cell aplasia in several patients owing to the formation of crossreactive neutralizing antibodies.

The application of rational engineering to immunogenicity has been aimed mostly at increasing the antigenicity of proteins for use in vaccines. Immune reduction of proteins as a whole is not as straightforward, and relatively few examples exist of rationally reducing or eliminating the immunogenicity of protein therapeutics. The only real success for immunogenicity reduction has been the humanization of murine antibodies, made possible by the high regularity of antibody sequence and structure and the ability to use proximity to human sequence as a metric for immunogenicity. In some cases, PEGylation has reduced the fraction of patients who raise neutralizing antibodies, possibly by sterically blocking access to epitopes [44]. Rational design methods that improve the solution properties of a protein therapeutic might also reduce immunogenicity because aggregates are generally more immunogenic than soluble proteins.

A more general approach to de-immunization involves mutagenesis of epitopes in the protein sequence and structure that are most responsible for stimulating the immune system. Some success has been achieved by randomly replacing surface residues, thus generating sequences with lower affinity for panels of known neutralizing antibodies [45,46]. An alternate approach is to disrupt T-cell activation by mutating peptides that bind class II major histocompatibility complex (MHC) alleles. Removal of MHC-binding epitopes offers a much more tractable approach to de-immunization than the removal of antibody epitopes because the diversity of MHC molecules comprises only 1-2x10^3 alleles, whereas the antibody repertoire is estimated to be approximately 10^8. A current challenge for rational design methods is to identify sequence variants that eliminate potential MHC-binding epitopes while maintaining protein structure and function. We anticipate that a general solution to the problem of protein immunogenicity will improve the safety and efficacy of protein therapeutics substantially and will enable new classes of nonhuman and *de novo* proteins to enter the clinic.

Antibodies

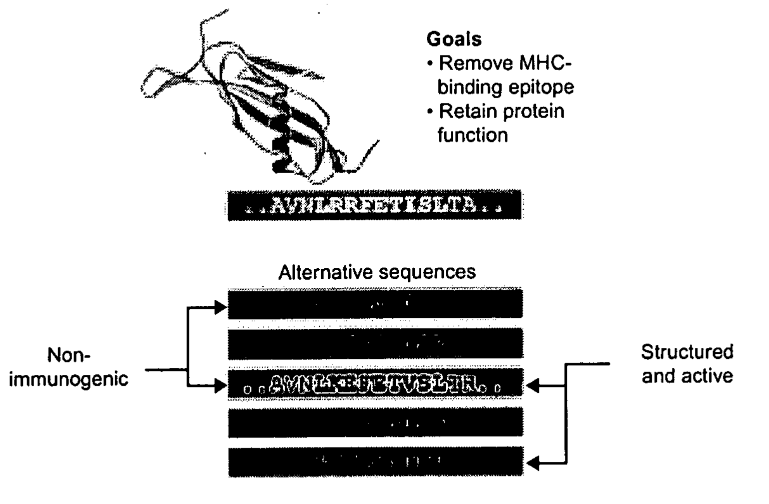
Some of the most visible and successful applications of rational engineering methods to biotherapeutics have occurred in the field of antibodies. Monoclonal antibodies are widely used as treatments for a variety of conditions from arthritis to cancer. There are currently 11 antibody products on the market, as shown in Table 3, and well over 100 in development. Despite such widespread acceptance and promise, there is still a need for structural and functional antibody optimization. Current antibody engineering efforts target both the variable and Fc regions of the molecule.

Antibody variable domains suffer from stability and solubility issues similar to all proteins, as discussed previously. However, because antibodies share a common structural scaffold, rational engineering studies have been able to dissect some of the sequence and structural determinants of variable region solubility and stability [47]. Notable developments include the structure-based design of more finely tuned complementarity determining region grafts and libraries [48,49], the use of phage-based selection methods for humanization [50,51] or fully human antibody generation [52] and the application of computational methods to increase the association rate of antibody/antigen formation at predicted 'ON-Rate AMPlification Sites' (Marvin and Lowman, pers. commun.). Furthermore, owing to the modular nature of immunoglobulin domains, variable domain architectures, such as diabodies, triabodies and bispecific diabodies, are being engineered to better serve specific therapeutic applications [53,54]. For more detail on variable region engineering the reader is referred to an excellent review by Maynard and Georgiou [53].

The Fc region of an antibody mediates interactions with several receptors, thus allowing antibodies to recruit the immune system and possess an extended serum half-life [55,56]. Significant effort has gone into engineering Fc for enhanced functional properties. Most exciting are recent results indicating that tighter binding by Fc to certain Fc gamma receptors, obtained by mutagenesis [57] or expression of carbohydrate isoforms [58,59], can result in enhanced effector function, potentially enabling the engineering of more potent antitumor antibodies. Additionally, some success has been achieved in modulating antibody PK by generating Fc variants with altered affinity for the neonatal receptor FcRn [60,61]. The bottleneck for Fc engineering is production. Because of the requirement for glycosylation, Fc and full-length antibodies must be produced in mammalian systems, precluding screening of large numbers of variants. Engineering a system with such high therapeutic potential yet limited screening capacity will be an exciting challenge for rational protein design.

State of the art rational engineering

The numerous examples discussed in this review illustrate both the demand for and power of rational engineering methods to improve the efficacy of biotherapeutics. There is currently an opportunity to replace the typical hit-and-miss approach to protein optimization with quantitative and systematic engineering strategies using computational



Drug Discovery Today

Figure 2. Multiparameter optimization of proteins. The modification of linear sequence epitopes is useful for modulating immunogenicity, altering proteolytic stability, introducing chemical modification sites and other strategies. However, these changes have a high likelihood of disrupting the integrity of the structure and function of the protein. It can be difficult to experimentally select for multiple properties simultaneously, such as nonimmunogenic and active variants. Rational design approaches are uniquely suited to the identification of protein sequences that satisfy multiple constraints. Abbreviation: MHC, major histocompatibility complex.

protein design methods [62–65]. Xencor's PDA™ technology is an example of these new methods (Box 1).

The full potential of computational design algorithms is realized when they are followed by high-throughput experimental screening efforts to single out superior members of a protein library. Computationally generated libraries are significantly enriched in stable, properly folded sequences relative to randomly generated libraries. In effect, structure-based sequence sampling methods yield an increased hit-rate, thereby decreasing the number of variants that must be screened. This feature is often critical to success because screens for therapeutic proteins, such as cell-based or *in vivo* assays, are often extremely low throughput. Given a high quality library, experimental screening methods can identify the sequence or sequences with the best characteristics quickly.

Computational design algorithms have tremendous potential for addressing conflicting constraints on a protein's sequence and structure, a common challenge in protein optimization efforts. As illustrated in Figure 2, many strategies (e.g. introduction of chemical or post-translational modification sites, removal of proteolysis sites and removal of MHC epitopes) require modifications to local primary structure. In most cases, however, the effect of these changes on the tertiary structure and functional integrity of the protein must also be considered. In other cases, one seeks primary structure alterations that disrupt one interaction while preserving a multiplicity of other interactions. Unfortunately, the number of acceptable sequence solutions narrows dramatically as the number of constraints is increased. One costly solution to this general problem is to develop assays that assess compatibility with each constraint separately. Alternatively, computational algorithms can simultaneously consider most or all of the constraints in the context of the whole protein. This approach also affords the opportunity to discover compensatory mutations elsewhere in the protein to accommodate changes made at the primary optimization site.

Conclusions

To convert a typical endogenous protein to a successful therapeutic it is often necessary to optimize several parameters, such as stability, solubility, PK and immunogenicity, while preserving or even enhancing function. Many strategies have already emerged for perturbing these parameters. We anticipate that the continued development and application of rational protein design technology will enable significant improvements in the efficacy and safety of existing protein therapeutics, as well as allow the generation of entirely novel classes of proteins and modes of action.

Acknowledgements

We would like to thank Bassil Dahiyat and David Szymkowski for helpful conversations and editorial advice.

References

1 Pharmafocus (2002) *Biotechnology 2006: Shaping the Future.* Available at: http://www.pharmafile.com/Pharmafocus/Features/feature.asp?fID=281

2 Voigt, C.A. *et al.* (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* 9, 553–558

3 Voigt, C.A. *et al.* (2001) Computationally focusing the directed evolution of proteins. *J. Cell. Biochem. Suppl.* 37, 58–63

4 Lin, L. (1998) Betaseron. *Dev. Biol. Stand.* 96, 97–104

5 Arakawa, T. *et al.* (1993) Cysteine 17 of recombinant human granulocyte colony stimulating factor is partially solvent-exposed. *J. Protein Chem.* 12, 525–531

6 Culajay, J.F. *et al.* (2000) Thermodynamic characterization of mutants of human fibroblast growth factor 1 with an increased physiological half-life. *Biochemistry* 39, 7153–7158

7 Desjarlais, J.R. and Handel, T.M. (1995) *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* 4, 2006–2018

8 Hellinga, H.W. and Richards, F.M. (1994) Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U. S. A.* 91, 5803–5807

9 Hurley, J.H. *et al.* (1992) Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.* 244, 1143–1159

10 Dahiyat, B.I. *et al.* (1997) Automated design of the surface positions of protein helices. *Protein Sci.* 6, 1333–1337

11 Malakauskas, S.M. and Mayo, S.L. (1998) Design, structure, and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* 5, 470–475

12 Marshall, S.A. *et al.* (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* 316, 189–199

13 Filikov, A.V. *et al.* (2002) Computational stabilization of human growth hormone. *Protein Sci.* 11, 1452–1461

14 Luo, P. *et al.* (2002) Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening. *Protein Sci.* 11, 1218–1226

15 Pipe, S.W. and Kaufman, R.J. (1997) Characterization of a genetically engineered inactivation-resistant coagulation factor VIIIa. *Proc. Natl. Acad. Sci. U. S. A.* 94, 11851–11856

16 Agren, L. *et al.* (1999) Hydrophobicity engineering of cholera toxin A1 subunit in the strong adjuvant fusion protein CTA1-DD. *Protein Eng.* 12, 173–178

17 Tan, P.H. *et al.* (1998) Engineering the isoelectric point of a renal cell carcinoma targeting antibody greatly enhances scFv solubility. *Immunotechnology* 4, 107–114

18 Sytkowski, A.J. *et al.* (1999) An erythropoietin fusion protein comprised of identical repeating domains exhibits enhanced biological properties. *J. Biol. Chem.* 274, 24773–24778

19 Goldenberg, M.M. (1999) Etanercept, a novel drug for the treatment of patients with severe, active rheumatoid arthritis. *Clin. Ther.* 21, 75–87

20 Syed, S. *et al.* (1997) Potent antithrombin activity and delayed clearance from the circulation characterize recombinant hirudin genetically fused to albumin. *Blood* 89, 3243–3252

21 Marques, J.A. *et al.* (2001) A barbourin-albumin fusion protein that is slowly cleared *in vivo* retains the ability to inhibit platelet aggregation *in vitro*. *Thromb. Haemost.* 86, 902–908

22 Dennis, M.S. *et al.* (2002) Albumin-binding as a general strategy of improving the pharmacokinetics of proteins. *J. Biol. Chem.* 277, 35035–35043

23 Vajo, Z. and Duckworth, W.C. (2000) Genetically engineered insulin analogs: diabetes in the new millenium. *Pharmacol. Rev.* 52, 1–9

24 Knauf, M.J. *et al.* (1988) Relationship of effective molecular size to systemic clearance in rats of recombinant interleukin-2 chemically modified with water-soluble polymers. *J. Biol. Chem.* 263, 15064–15070

25 Clark, R. *et al.* (1996) Long-acting growth hormones produced by conjugation with polyethylene glycol. *J. Biol. Chem.* 271, 21969–21977

26 Harris, J.M. *et al.* (2001) Pegylation: a novel process for modifying pharmacokinetics. *Clin. Pharmacokinet.* 40, 539–551

27 Roberts, M.J. *et al.* (2002) Chemistry for peptide and protein PEGylation. *Adv. Drug Deliv. Rev.* 54, 459–476

28 Kinstler, O. *et al.* (2002) Mono-N-terminal poly(ethylene glycol)-protein conjugates. *Adv. Drug Deliv. Rev.* 54, 447–485

29 Bailon, P. *et al.* (2001) Rational design of a potent, long-lasting form of interferon: a 40-kDa-branched polyethylene glycol-conjugated interferon alpha-2a for the treatment of hepatitis C. *Bioconjug. Chem.* 12, 195–202

30 Wang, Y.S. *et al.* (2002) Structural and biological characterization of pegylated recombinant interferon alpha-2b and its therapeutic implications. *Adv. Drug Deliv. Rev.* 54, 547–570

31 MacDougall, I.C. *et al.* (1999) Pharmacokinetics of novel erythropoiesis stimulating protein compared with epoetin alfa in dialysis patients. *J. Am. Soc. Nephrol.* 10, 2392–2395

32 French, A.R. *et al.* (1995) Intracellular traffiking of epidermal growth factor family ligands is directly influenced by the pH sensitivity of the receptor/ligand interaction. *J. Biol. Chem.* 270, 4334–4340

33 Sarkar, C.A. *et al.* (2002) Rational cytokine design for increased lifetime and enhanced potency using pH activated histidine-switching. *Nat. Biotechnol.* 20, 908–913

34 Weintraub, B.D. and Szkudlinski, M.W. (1999) Development and *in vitro* characterization of human recombinant thyrotropin. *Thyroid* 9, 447–450

35 Grossmann, M. *et al.* (1998) A rational design strategy for protein hormone superagonists. *Nat. Biotechnol.* 16, 871–875

36 Siemeister, G. *et al.* (1998) An antagonistic vascular endothelial grown factor (VEGF) variant inhibits VEGF-stimulated receptor autophosphorylation and proliferation of human endothelial cells. *Proc. Natl. Acad. Sci. U. S. A.* 95, 4625–4669

37 Savino, R. *et al.* (1994) Rational design of a receptor super-antagonist of human interleukin-6. *EMBO J.* 13, 5863–5870

38 Fuh, C. *et al.* (1992) Rational design of potent antagonists to the human growth hormone receptor. *Science* 256, 1677–1680

39 Olson, K. *et al.* (1998) *Human Growth Hormone Variants.* US patent number 5849535

40 Lu, C. *et al.* (2001) An isolated, surface expressed I domain of the integrin alphaL beta2 is sufficient for strong adhesive function when locked in the open conformation with a disulfide bond. *Proc. Natl. Acad. Sci. U. S. A.* 98, 2387–2392

41 Shimaoka, M. *et al.* (2001) Reversibly locking a protein fold in an active conformation with a disulfide bond: integrin alphaL I domains with high affinity and antagonist activity *in vivo. Proc. Natl. Acad. Sci. U. S. A.* 98, 6009–6014

42 Shimaoka, M. *et al.* (2000) Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* 7, 674–678

43 Antonelli, G. and Dianzani, F. (1999) Development of antibodies to interferon beta in patients: technical and biological aspects. *Eur. Cytokine Netw.* 10, 413–422

44 He, X.H. *et al.* (1999) Reducing the immunogenicity and improving the *in vivo* activity of trichosanthin by site-directed pegylation. *Life Sci.* 65, 355–368

45 Meyer, D.L. *et al.* (2001) Reduced antibody response to streptavidin through site-directed mutagenesis. *Protein Sci.* 10, 491–503

46 Laroche, Y. *et al.* (2000) Recombinant staphylokinase variants with reduced antigenicity due to elimination of B-lymphocyte epitopes. *Blood* 96, 1425–1432

47 Worn, A. and Pluckthun, A. (2001) Stability engineering of antibody single chain Fv fragments. *J. Mol. Biol.* 305, 989–1010

48 Tamura, M. *et al.* (2000) Structural correlates of an anticarcinoma antibody: identification of specificity-determining residues (SDRs) and development of a minimally immunogenic antibody variant by retention of SDRs only. *J. Immunol.* 164, 1432–1441

49 Chen, Y. *et al.* (1999) Selection and analysis of an optimized anti-VEGF antibody: crystal structure of an affinity-matured Fab in complex with antigen. *J. Mol. Biol.* 293, 865–881

50 Rader, C. *et al.* (1998) A phage display approach for rapid antibody humanization: designed combinatorial V gene libraries. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8910–8915

51 Baca, M. *et al.* (1997) Antibody humanization using monovalent phage display. *J. Biol. Chem.* 272, 10678–10684

52 Griffiths, A.D. and Duncan, A.R. (1998) Strategies for selection of antibodies by phage display. *Curr. Opin. Biotechnol.* 9, 102–108

53 Maynard, J. and Georgiou, G. (2000) Antibody engineering. *Annu. Rev. Biomed. Eng.* 2, 339–376

54 Pluckthun, A. and Pack, P. (1997) New protein engineering approaches to multivalent and bispecific antibody fragments. *Immunotechnology* 3, 83–105

55 Ravetch, J.V. and Bolland, S. (2001) IgG Fc receptors. *Annu. Rev. Immunol.* 19, 275–290

56 Raghavan, M. and Bjorkman, P.J. (1996) Fc receptors and their interactions with immunoglobulins. *Annu. Rev. Cell Dev. Biol.* 12, 181–220

57 Shields, R.L. *et al.* (2001) High resolution mapping of the binding site on human IgG1 for Fc gamma RI, Fc gamma RII, Fc gamma RIII, and FcRn and design of IgG1 variants with improved binding to the Fc gamma R. *J. Biol. Chem.* 276, 6591–6604

58 Shields, R.L. *et al.* (2002) Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. *J. Biol. Chem.* 277, 26733–26740

59 Umana, P. *et al.* (1999) Engineered glycoforms of an antineuroblastoma IgG1 with optimized antibody-dependent cellular cytotoxic activity. *Nat. Biotechnol.* 17, 176–180

60 Ghetie, V. *et al.* (1997) Increasing the serum persistence of an IgG fragment by random mutagenesis. *Nat. Biotechnol.* 15, 637–640

61 Kim, J.K. *et al.* (1999) Mapping the site on human IgG for binding of the MHC class I-related receptor, FcRn. *Eur. J. Immunol.* 29, 2819–2825

62 Street, A.G. and Mayo, S.L. (1999) Computational protein design. *Structure Fold Des.* 7, R105–109

63 Mendes, J. *et al.* (2002) Energy estimation in protein design. *Curr. Opin. Struct. Biol.* 12, 441–446

64 Pokala, N. and Handel, T.M. (2001) Review: Protein design- Where we were, where we are, where we're going. *J. Struct. Biol.* 134, 269–281

65 Kraemer-Pecore, C.M. *et al.* (2001) Computational protein design. *Curr. Opin. Chem. Biol.* 5, 690–695

66 Dahiyat, B.I. and Mayo, S.L. (1996) Protein design automation. *Protein Sci.* 5, 895–903

67 Dahiyat, B.I. and Mayo, S.L. (1997) *De novo* protein design: fully automated sequence selection. *Science* 278, 82–87

Reprint Series
3 October 1997, Volume 278, pp. 80–81

SCIENCE

Proteins from Scratch

William F. DeGrado

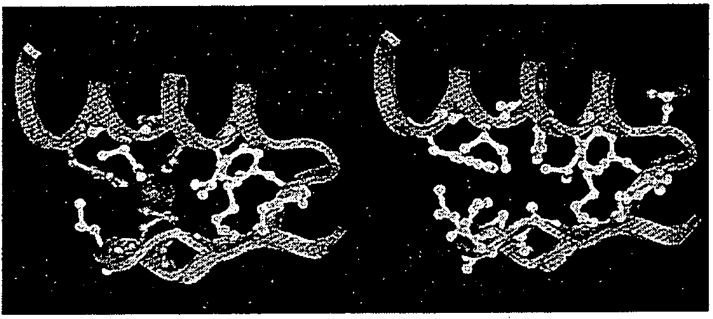
Proteins from Scratch

William F. DeGrado

Not long ago, it seemed inconceivable that proteins could be designed from scratch. Because each protein sequence has an astronomical number of potential conformations, it appeared that only an experimentalist with the evolutionary life span of Mother Nature could design a sequence capable of folding into a single, well-defined three-dimensional structure. But now, on page 82 of this issue, Dahiyat and Mayo (1) describe a new approach that makes de novo protein design as easy as running a computer program. Well almost. . .

The intellectual roots of this new work go back to the early 1980s when protein engineers first thought about designing proteins (2). At that point, the prediction of a protein's three-dimensional structure from its sequence alone seemed a difficult proposition. However, they opined that the inverse problem—designing an amino acid sequence capable of assuming a desired three-dimensional structure—would be a more tractable problem, because one could "over-engineer" the system to favor the desired folding pattern. Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights, and paradigms gleaned from two decades of experiments (3) into a single computational algorithm that predicts an optimal sequence for a given fold. Further, when put to the test the algorithm actually predicted a sequence that folded into the desired three-dimensional structure. Thus, the rules of protein folding and computational methods for de novo design may now be sufficiently defined to allow the engineering of a variety of proteins.

Dahiyat and Mayo's program divides the interactions that stabilize protein structures into three categories: interactions of side chains that are exposed to solvent, of side chains buried in the protein interior, and of parts of the protein that occupy more interfacial positions. Exposed residues contribute to stability, primarily through conformational preferences and weakly attractive, solvent-exposed polar interactions (4). The burial of hydrophobic residues in the well-packed interior of a protein provides an even more powerful driving force for folding. The side chains in the interior of a protein adopt unique conformations, the prediction of which is a large combinatorial problem.



Better than the real thing. The natural zinc finger protein Zif268 (**left**) is stabilized in part by a core of hydrophobic (green) side chains and metal-chelating side chains (red). In the designed protein FSD-1 (**right**), the Zif268 core is retained but the metal-chelating His residues and one of the Cys residues of Zif268 are converted to hydrophobic Phe and Ala residues, thereby extending the hydrophobic core. The fourth metal ligand Cys^8 is converted to a Lys residue. The apolar portion of this interfacial residue shields the hydrophobic core, whereas its ammonium group is exposed to solvent. The helix is also stabilized by an N-capping interaction (19), which presumably also stabilizes the structure.

One important simplifying assumption arose from the early work of Jainin *et al.* (5), who showed that each individual side chain can adopt a limited number of low-energy conformations (named rotamers), reducing the number of probable conformers available to a protein. This work was subsequently extended to the design of proteins containing only the most favorable rotamers (6). Although the side chains in natural proteins deviate from ideality in a few cases (complicating the prediction of the structures of natural proteins), these deviations need not be considered in the design of idealized proteins. Thus, various algorithms have been developed to examine all possible hydrophobic residues in all possible rotameric states, to find combinations that efficiently fill the interior of a protein. A complementary approach uses genetic methods to exhaustively search for sequences capable of filling a protein core (7), and this work has been adapted for the de novo design of proteins (8).

Interfacial residues are also quite important for protein stability (9, 10). They are often amphiphilic (for example, Lys, Arg, and Tyr) and their apolar atoms can cap the hydrophobic core, while their polar groups engage in electrostatic and hydrogen-bonded interactions.

Until recently, protein designers have frequently concentrated on quantifying the energetics associated with just one of these three types of interactions (3). However, de novo design is best approached by simultaneously considering all of the side chains in the protein—unfortunately, a very high-order combinatorial problem. For instance, the volume available to the interior side chains depends on the nature and conformation of the residues at the interfacial positions and vice versa. Dahiyat and Mayo assumed that each of these three features had been adequately quantitated to provide a useful empirical energy function for protein design. Their program combines a number of feaures taken from earlier potential functions and includes a penalty for exposing hydrophobic groups to solvent. Another essential innovation included in their program is an implementation of the Dead-End Elimination theorem, to efficiently search through sequence and side chain rotamer space.

Dahiyat and Mayo's target fold is a zinc finger, a motif with a well-established history in protein structure prediction and design. In an early, prescient paper, Berg correctly inferred that this His^{2}Cys^2 Zn-binding motif must feature a β-β-α fold that would position the ligating groups in a tetrahedral array around the bound Zn(II) (11). Favorable metal ion–ligand interactions together with a small apolar core help stabilize the three-dimensional structure of this compact fold. More recently, Imperiali and co-workers have designed a peptide that folded into this motif, even in the absence of metal ions (12). The design included a D-amino acid to stabilize a type II′ turn, and a large, rigid tricyclic side chain that may help consolidate the hydrophobic core. This work was particularly ex-

An enhanced version of this Perspective with links to additional resources is available for *Science* Online subscribers at www.sciencemag.org

The author is in the Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6059, USA. E-mail: wdegrado@mail.med.upenn.edu

citing because, before their studies, it was not expected that sequences as short as 25 residues in length could fold into stable tertiary structures.

Now, Dahiyat and Mayo take these studies one step further through the design of a sequence composed of only natural amino acids that adopts the zinc finger motif. As input to their program, they introduced the coordinates of the backbone atoms from the crystal structure of the second domain of the zinc finger protein Zif268. The program then evaluated a total of 10^{62} possible side chain–rotamer combinations to find a sequence capable of stabilizing this fold without a bound metal ion. The resulting protein sequence shares a small hydrophobic core with its predecessor from Zif268. However, in the newly designed protein FSD-1 the core is enlarged through the addition of hydrophobic residues that fill the space vacated by the removal of the metal-binding site (see the figure). This increase in the size of the hydrophobic core together with the enhancements in the propensity for forming the appropriate secondary structure provide an adequate driving force for folding. The designed miniprotein actually folds into the desired structure as assessed by nuclear magnetic resonance spectroscopy, and the observed structure closely resembles the three-dimensional structure of Zif268.

Because of its small size, the protein is marginally stable. A Van't Hoff analysis of the thermal unfolding curve gives a change in the enthalpy (ΔH_{vH}) of approximately –10 kcal/mol, and indicates that the protein is about 90 to 95% folded at low temperatures (*13*). The small value ΔH_{vH} and the lack of strong cooperativity in the unfolding transition are expected for a native-like protein of this very small size (*14*). Thus, FSD-1 is the smallest protein known to be capable of folding into a unique structure without the thermodynamic assistance of disulfides, metal ions, or other subunits. This important accomplishment illustrates the impressive ability of Dahiyat and Mayo's program to design highly optimized sequences.

This new achievement caps a banner year for de novo protein design. Earlier, Regan (*15*) answered the challenge of changing a protein's tertiary structure by altering no more than 50% of its sequence. And although Dahiyat and Mayo have demonstrated that the stabilizing metal-binding site is not necessary in their system, Caradonna, Hellinga, and co-workers (*16*) have made impressive progress in automating the introduction of functional metal-binding sites into the three-dimensional structures of natural proteins. Further, other workers (*17*) have used less automated approaches to successfully introduce functionally and spectroscopically interesting metal-binding sites into de novo designed proteins.

To date, the most computationally intensive protein design problems have been the redesign of natural proteins of known three-dimensional structure. But the new automated approaches open the door to the de novo design of structures with entirely novel backbone conformations. It will be interesting to see if Dahiyat and Mayo's approach of designing an optimal sequence for a given fold is sufficient, or if it will be necessary also to destabilize alternate possible folds. Indeed, when using an earlier version of their algorithm to repack the interior of the coiled coil from GCN4, they had to retain the identity of a buried Asn residue from the wild-type protein. Although the inclusion of this Asn actually destabilized the desired fold, it was nevertheless essential to avoid the formation of alternate, unwanted conformers (*18*). The ability to ask such focused questions will reveal much about how natural proteins adopt their folded conformations while simultaneously allowing the design of entirely new polymers for applications ranging from catalysis to pharmaceuticals.

References and Notes

1. B. I. Dahiyat and S. L. Mayo, *Science* **278**, 82.
2. K. E. Drexler, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5275 (1981); C. Pabo, *Nature* **301**, 200 (1983).
3. W. F. DeGrado, Z. R. Wasserman, J. D. Lear, *Science* **243**, 622 (1989); J. W. Bryson *et al.*, *ibid.* **270**, 935 (1995); M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Opin. Struct. Biol.* **6**, 3 (1996).
4. R. Munoz and L. Serrano, *Proteins* **20**, 301 (1994); C. A. Kim and J. M. Berg, *Nature* **362**, 267 (1993); D. L. Minor and P. S. Kim, *ibid.* **367**, 660 (1994); C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510 (1994).
5. J. Janin, S. Wodak, M. Levitt, B. Maigret, *J. Mol. Biol.* **125**, 37 (1978).
6. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987); J. R. Desjarlais and T. M. Handel, *Protein Sci.* **4**, 2006 (1995); X. Jing, E. J. Bishop, R. S. Farid, *J. Am. Chem. Soc.* **119**, 838 (1997).
7. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
8. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *ibid.* **262**, 1680 (1993).
9. K. J. Lumb and P. S. Kim, *ibid.* **271**, 1137 (1996); Y. Yu, O. D. Monera, R. S. Hodges, P. L. Privalov, *J. Mol. Biol.* **255**, 367, (1996).
10. A. C. Braisted and J. A. Wells, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5688 (1996).
11. J. M. Berg, *ibid.* **85**, 99 (1988).
12. M. D. Struthers, R. P. Cheng, B. Imperiali, *Science* **271**, 342 (1996).
13. This Van't Hoff analysis of the protein is approximate because of the lack of definition of the pre- and posttransition baselines.
14. P. Alexander, S. Fahnestock, T. Lee, J. Orban, P. Bryn, *Biochemistry* **31**, 3597 (1992).
15. S. Dalal, S. Balasubramanian, L. Regan, *Nat. Struct. Biol.* **4**, 548 (1997).
16. A Pinto, H. W. Hellinga, J. P. Caradonna, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5562 (1997); C. Coldren, H. W. Hellinga, J. P. Caradonna, *ibid.*, p. 6635.
17. B. R. Gibney, S. E. Mulholland, F. Rabanal, P. L. Dutton, *ibid.* **93**, 15041 (1996); M. P. Scott, J. Biggins, *Protein Sci.* **6**, 340 (1997); P. A. Arnold, W. R. Shelton, D. R. Benson, *J. Am. Chem. Soc.* **119**, 3181 (1997); G. R. Dieckman *et al.*, *ibid.*, p. 6195.
18. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401 (1993); K. J. Lumb and P. S. Kim, *Biochemistry* **34**, 8642 (1995).
19. L. G. Presta and G. D. Rose, *Science* **240**, 1632 (1988); J. S. Richardson and D. C. Richardson, *ibid.*, p. 1648.

US006627186B1

(12) **United States Patent**
Dahiyat et al.

(10) **Patent No.: US 6,627,186 B1**
(45) **Date of Patent: Sep. 30, 2003**

(54) **NUCLEIC ACIDS AND PROTEIN VARIANTS OF HG-CSF WITH GRANULOPOIETIC ACTIVITY**

(75) Inventors: **Bassil I. Dahiyat**, Los Angeles, CA (US); **Peizhi Luo**, Arcadia, CA (US)

(73) Assignee: **Xencor**, Monrovia, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/479,313**

(22) Filed: **Jan. 6, 2000**

Related U.S. Application Data

(60) Provisional application No. 60/115,131, filed on Jan. 6, 1999, and provisional application No. 60/118,831, filed on Feb. 5, 1999.

(51) **Int. Cl.**^{7} **A61K 38/19**; C07K 14/535; C12N 1/21; C12N 5/10; C12N 15/27

(52) **U.S. Cl.** **424/85.1**; 514/2; 514/12; 435/69.1; 435/320.1; 435/252.3; 435/254.11; 435/325; 536/23.1; 536/23.5; 530/351; 530/350

(58) **Field of Search** 530/350, 351, 530/399; 435/69.1, 325, 365.1, 320.1, 252.3, 254.11; 536/23.5, 23.1; 514/2, 8; 424/85.1

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,810,643	A		3/1989	Souza
4,833,127	A		5/1989	Ono et al.
4,999,291	A		3/1991	Souza
5,214,132	A		5/1993	Kuga et al.
5,218,092	A	*	6/1993	Sasaki et al.
5,362,853	A		11/1994	Kuga et al.
5,399,345	A	*	3/1995	Schumacher et al.
5,416,195	A		5/1995	Camble et al.
5,580,755	A		12/1996	Souza
5,581,476	A		12/1996	Osslund
5,790,421	A		8/1998	Osslund
5,830,705	A		11/1998	Souza

FOREIGN PATENT DOCUMENTS

EP	0 459 630	4/1991
WO	94/17185	8/1994
WO	98/47089	10/1998

OTHER PUBLICATIONS

Kuwabara et al. 1992, J. Pharamacobio–Dyn. vol. 15: pp. 121–129, Highly sensitive enzyme–linked immunosorbent assay for marograstim (KW–2228), a mutant of human granulocyte stimulating factor.*

Dahiyat et al., "Protein design automation," Protein Science, 5:895–903 (1996).

Kuga et al., "Mutagenesis of human granulocyte colony stimulating factor," Biochemical and Biophysical Research Communications, 159(1):103–111 (1989).

Luo et al., "Automated Design of Enhanced Granulopoietic Proteins," FASEB Journal, 13(7): a1431 (1999).

* cited by examiner

Primary Examiner—Elizabeth Kemmerer
(74) *Attorney, Agent, or Firm*—Dorsey & Whitney LLP; Robin M. Silva; Renee M. Kosslak

(57) **ABSTRACT**

The invention relates to novel granulopoietic activity (GPA) proteins and nucleic acids. The invention further relates to the use of the GPA proteins in the treatment of G-CSF related disorders.

27 Claims, 16 Drawing Sheets

```
TPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWA
PLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIW
QQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP
```

hGCSF wild type

```
ATGACTCCATTAGGTCCAGCTTCCTCTCTGCCGCAAAGCTTCCTGCTGAAATGCCTGGAACAGGTTCGTAAAATCCAGGGTGATGG
TGCTGCTCTGCAGGAAAAACTGTGCGCTACCTACAAACTGTGCCATCCGGAAGAACTGGTTCTGCTGGGTCACTCCCTGGGTATCC
CGTGGGCGCCGCTGAGCTCCTGCCCGAGCCAGGCTCTGCAGCTGGCTGGTTGCCTGTCCCAATTGCACAGCGGCCTTTTCCTGTAC
CAGGGTCTGCTGCAAGCTCTGGAAGGTATCTCCCCGGAACTGGGTCCGACCCTGGACACTCTGCAGCTGGACGTCGCTGACTTCGC
TACCACCATCTGGCAGCAGATGGAAGAACTGGGTATGGCTCCGGCTCTGCAGCCGACCCAGGGTGCTATGCCGGCTTTCGCTTCCG
CTTTCCAGCGTCGCGCAGGTGGCGTTCTGGTTGCTAGCCACCTGCAGAGCTTCCTGGAAGTTTCCTACCGTGTTCTGCGTCACCTG
GCTCAGCCGTGA
```

FIG._1

core3	17	21	24	28	31	35	41	47	54	56	75	78	82	85	88	89	92
	95	99	103	106	110	113	114	117	140	151	152	153	154	157	160	161	168
core4	17	21	24	28	31	75	78	82	85	89	103	106	110	113	114	117	149
	150	151	152	153	154	157	160	161	168								
core4v	17	21	24	28	31	75	78	82	85	89	103	106	110	113	114	117	151
	152	153	154	157	160	161	167	168									
bndry4_2		14	20	27	32	34	38	77	79	84	91	99	102	107	109	116	
		120	145	146	147	148	155	156	164	170							
bndry4_core4		14	20	27	32	34	38	77	79	84	91	99	102	107	109	116	
		120	145	146	147	148	155	156	164	170							
bndry4_AD		14	20	27	32	34	38	145	146	147	148	155	156	164	170		
bndry4_AD_core4		14	20	27	32	34	38	145	146	147	148	155	156	164	170		

FIG._2

G-CSF Designs - Optimal Sequences Selected by PDA*

```
                   1        10        20        30        40        50        60
hGCSFwt           MTPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAP
bndry4_2                        I     L      E    I I   K
bndry4_core4                    I  L  L      EA   L E   H
bndry4_AD                       I     L      E    I E   H
bndry4_AD_core4                 I  L  L      EA   L E   H
core4                              L          A
core4_V167A                        L   I      A
core3                              L          A      I
--------------------------------------------------------------------------------
sm0                                A          A
fm2                                A          A
fm3                                L          A
fm4                                L          A
fm7                                L          A

                           70        80        90        100       110       120
hGCSFwt           LSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQ
bndry4_2                            L           K       V  I    I E      I   L
bndry4_core4                       FL     F     K          KV   I EI  L  I   L
bndry4_AD
bndry4_AD_core4                    F      F                 V     I  L
core4                              F      F                 V     I  L
core4_V167A                        F      F                       I  L
core3                              F      F   F  F                   L
--------------------------------------------------------------------------------
sm0
fm2
fm3                                F                                 L
fm4                                F                                 L
fm7                                F      F                          L
```

FIG._3A

```
                          130       140       150       160       170
hGCSFwt          MEELGMAPALQPTQGAMPAFASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP
bndry4_2                                  KED      IL       A
bndry4_core4                              KED  I I IL       A   F
bndry4_AD                                 KET      IL       A
bndry4_AD_core4                           KED  I I IL       A   F
core4                                          I I              F
cor 4_V167A                                    I I   I  WF     AF
cor 3                                          I I              F
-----------------------------------------------------------------------------------
sm0
fm2                                            I I
fm3                                                             F
fm4                                            I I              F
fm7                                            I I              F
```

*Sequences shown below dotted lines were not obtained from PDA calculations but were d rived by reverting some core4 or core3 mutant positions to wild type. Core4 mutant positions are indicated in bold. The sequence selected for Core4_V167A is not the ground state; Monte Carlo analysis shows the ground state with Phe instead of Trp for position 160, and Leu instead of Phe for position 161 (see Table 4).

FIG._3B

Core4 - Monte Carlo Analysis - Ground State and Allowed Amino Acids and Their Number of Occurrences (For theTop 1000 Sequences)

hG-CSF	Position	Ground State					
CYS	17	LEU	736	ILE	229		
VAL	21	VAL	687	ILE	287		
ILE	24	VAL	38	ILE	961		
GLY	28	ALA	747	LEU	172		
LEU	31	VAL	251	LEU	707		
LEU	75*	LEU	999				
LEU	78	PHE	974				
LEU	82*	LEU	974				
TYR	85	PHE	847	TYR	140		
LEU	89	LEU	628	PHE	321		
LEU	103	VAL	351	LEU	264	ILE	313
LEU	106*	LEU	940				
VAL	110	VAL	415	LEU	143	ILE	441
PHE	113	LEU	999				
ALA	114*	ALA	999				
ILE	117*	ILE	956				
GLY	149*	GLY	999				
GLY	150*	GLY	999				
VAL	151	ILE	999				
LEU	152*	LEU	999				
VAL	153	VAL	411	ILE	588		
ALA	154*	ALA	999				
LEU	157	LEU	805	ILE	187		
PHE	160	PHE	565	TRP	434		
LEU	161	LEU	838	PHE	161		
LEU	168	PHE	999				

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._4

Table 4. **Core4v - Monte Carlo Analysis** (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

hG-CSF	Position	Ground State							
CYS	17	LEU	697	VAL	51	ILE	251		
VAL	21	VAL	682	VAL	682	ILE	300	PHE	17
ILE	24	ILE	938	VAL	61				
GLY	28	ALA	806	LEU	193				
LEU	31	LEU	694	ALA	1	VAL	257	ILE	47
LEU	75*	LEU	999						
LEU	78	PHE	982	VAL	17				
LEU	82	LEU	982	PHE	17				
TYR	85	PHE	887	VAL	2	ILE	16	TYR	94
LEU	89	LEU	637	PHE	314	TRP	48		
LEU	103	VAL	357	ALA	78	LEU	269	ILE	295
LEU	106	LEU	945	VAL	54				
VAL	110	ILE	445	VAL	405	LEU	149		
PHE	113	LEU	999						
ALA	114*	ALA	999						
ILE	117*	ILE	938	VAL	61				
VAL	151	ILE	999						
LEU	152*	LEU	999						
VAL	153	ILE	585	VAL	414				
ALA	154*	ALA	999						
LEU	157	LEU	797	VAL	18	ILE	184		
PHE	160	PHE	551	TRP	448				
LEU	161	LEU	843	PHE	156				
VAL	167	ALA	999						
LEU	168	PHE	999						

*position where Monte Carlo didn't find an altemative and where the top amino acid is the wild type

FIG._5

Table 5. **Core3 - Monte Carlo Analysis** (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

hG-CSF	Position	Ground State											
CYS	17	LEU	585	VAL	35	ILE	379						
VAL	21	VAL	551	ALA	15	ILE	291	PHE	141	TYR	1		
ILE	24	ILE	657	ALA	31	VAL	303	LEU	8				
GLY	28	ALA	928	LEU	71								
LEU	31	LEU	888	VAL	111								
LYS	35	ILE	785	VAL	214								
LEU	41*	LEU	999										
LEU	47*	LEU	999										
LEU	54*	LEU	999										
ILE	56*	ILE	999										
LEU	75*	LEU	999										
LEU	78	PHE	692	ALA	10	VAL	149	LEU	12	ILE	75	TYR	61
LEU	82	LEU	851	ALA	12	PHE	136						
TYR	85	PHE	636	TRP	363								
LEU	88*	LEU	999										
LEU	89	PHE	674	LEU	214	TRP	111						
LEU	92	PHE	999										
ILE	95*	ILE	999										
LEU	99*	LEU	999										
LEU	103	LEU	888	ILE	111								
LEU	106	LEU	893	VAL	106								
VAL	110	VAL	400	ALA	14	LEU	294	ILE	291				
PHE	113	LEU	954	ALA	1	PHE	44						
ALA	114*	ALA	999										
ILE	117	ILE	790	ALA	15	VAL	168	LEU	5	PHE	20	TRP	1

FIG._6A

Table 5. Core3 - Monte Carlo Analysis (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

hG-CSF	Position	Ground State							
PHE	140*	PHE	999						
VAL	151	ILE	999						
LEU	152*	LEU	999						
VAL	153	ILE	999						
ALA	154*	ALA	999						
LEU	157	LEU	694	ALA	22	VAL	179	ILE	104
PHE	160	PHE	574	TRP	425				
LEU	161	LEU	784	ALA	6	VAL	55	PHE	154
LEU	168	PHE	999						

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._6B

Table 6. Bndry4_2 - Monte Carlo Analysis (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

hG-CSF	Position	Ground State							
LEU	14	ILE	998						
GLN	20	LEU	999						
ASP	27	GLU	999						
GLN	32	ILE	999						
LYS	34	LYS	717	ILE	209	GLU	73		
THR	38	VAL	409	ILE	188	GLU	237	LYS	154
GLN	77*	GLN	999						
HIS	79	LEU	999						
LEU	84*	LEU	999						
ALA	91	LYS	999						
LEU	99	VAL	759	LEU	193				
THR	102	LEU	562	ILE	404				
GLN	107	ILE	993						
VAL	109	GLU	525	VAL	474				
THR	116	ILE	749	LEU	198	LYS	52		
GLN	120	LEU	999						
GLN	145	GLN	650	GLU	349				
ARG	146	LYS	891	GLN	108				
ARG	147	GLU	999						
ALA	148	THR	401	ALA	268	ASP	330		
SER	155	ILE	999						
HIS	156	LEU	999						
SER	164	ALA	999						
HIS	170	HSP	380	LEU	111	GLU	248	GLN	227

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._7

Table 7. Bndry4_core4 - Monte Carlo Analysis (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

hG-CSF	Position	Ground State						
LEU	14	ILE 941	LEU 58	ILE 941				
GLN	20	LEU 999						
ASP	27	GLU 970	SER 29					
GLN	32	LEU 631	VAL 125		ILE 243			
LYS	34	GLU 961	GLN 22	LYS 16				
THR	38	HSP 931	VAL 19	ILE 4	GLU 5	LYS 40		
GLN	77*	GLN 999						
HIS	79	LEU 999						
LEU	84*	LEU 999						
ALA	91	LYS 999						
LEU	99	LEU 922	GLU 77					
THR	102	LYS 729	THR 14	VAL 150	LEU 2	ILE 71	GLU 14	GLN 19
GLN	107	ILE 968	VAL 30	LEU 1				
VAL	109	GLU 591	VAL 402	ASP 2	GLN 4			
THR	116	ILE 647	VAL 15	LEU 275	GLU 1	LYS 61		
GLN	120	LEU 999						
GLN	145	GLN 658	GLU 341					
ARG	146	LYS 857	GLN 142					
ARG	147	GLU 998	LYS 1					
ALA	148	ASP 359	ALA 310	THR 330				
SER	155	ILE 999						
HIS	156	LEU 999						
SER	164	ALA 999						
HIS	170	HSP 380	ASP 26	LEU 109	GLU 239	GLN 214	LYS 31	

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._8

Table 8. Bndry4_AD - Monte Carlo Analysis (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

Position	Ground State											
14	ILE	887	LEU	112								
20	LEU	999										
27	GLU	984										
32	ILE	931										
34	GLU	357	ILE	68	GLN	223	LYS	277				
38	VAL	287	ILE	133	HSP	225	GLU	217	LYS	123		
145	GLN	605	GLU	394								
146	LYS	786	GLN	213								
147	GLU	962										
148	THR	373	ALA	305	ASP	321						
155	ILE	976										
156	LEU	994										
164	ALA	999										
170	HSP	304	ASP	55	LEU	136	GLU	230	GLN	209	LYS	62

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._9

Table 9. Bndry4_AD_core4 - Monte Carlo Analysis (Ground State and Allowed Amino Acids and Their Number of Occurrences (For the Top 1000 Sequences)

Position	Ground State									
14	ILE	896	LEU	103						
20	LEU	999								
27	GLU	996								
32	LEU	523	VAL	194	ILE	271				
34	GLU	400	GLN	207	LYS	341				
38	VAL	300	ILE	89	HSP	277	GLU	203	LYS	130
145	GLN	623	GLU	376						
146	LYS	820	GLN	179						
147	GLU	986								
148	ASP	344	ALA	332	THR	323				
155	ILE	998								
156	LEU	996								
164	ALA	999								
170	HSP	330	LEU	134	GLU	234	GLN	216		

*position where Monte Carlo didn't find an alternative and where the top amino acid is the wild type

FIG._10

Core3
ATGACTCCATTAGGTCCAGCTTCCTCTCTGCCGCAAAGCTTCCTGCTGAAACTGCTGGAACAGGTTCGTAAAATCCAGGGTGATGC
AGCTGCTCTGCAGGAAAAAATCTGCGCTACCTACAAACTGTGCCATCCGGAAGAACTGGTTCTGCTGGGTCACTCCCTGGGTATCC
CGTGGGCGCCGCTGAGCTCCTGCCCGAGCCAGGCTCTGCAGCTGGCTGGTTGCCTGTCCCAATTCCACAGCGGCCTTTTCCTGTTC
CAGGGTCTGTTCCAGGCTTTCGAAGGTATCTCCCCGGAACTGGGTCCGACCCTGGACACTCTGCAGCTGGACGTCGCTGACCTGGC
TACCACCATCTGGCAGCAGATGGAAGAACTGGGTATGGCTCCGGCTCTGCAGCCGACCCAGGGTGCTATGCCGGCTTTCGCTTCCG
CTTTCCAGCGTCGCGCAGGTGGCATCCTGATCGCTAGCCACCTGCAGAGCTTCCTGGAAGTTTCCTACCGTGTTTTCCGTCACCTG
GCTCAGCCGTGA

FIG._11A

Core4
ATGACTCCATTAGGTCCAGCTTCCTCTCTGCCGCAAAGCTTCCTGCTGAAACTGCTGGAACAGGTTCGTAAAATCCAGGGTGATGC
AGCTGCTCTGCAGGAAAAACTGTGCGCTACCTACAAACTGTGCCATCCGGAAGAACTGGTTCTGCTGGGTCACTCCCTGGGTATCC
CGTGGGCGCCGCTGAGCTCCTGCCCGAGCCAGGCTCTGCAGCTGGCTGGTTGCCTGTCCCAATTCCACAGCGGCCTTTTCCTGTTC
CAGGGTCTGCTGCAAGCTCTGGAAGGTATCTCCCCGGAACTGGGTCCGACCGTTGACACTCTGCAGCTGGACATCGCTGACCTGGC
TACCACCATCTGGCAGCAGATGGAAGAACTGGGTATGGCTCCGGCTCTGCAGCCGACCCAGGGTGCTATGCCGGCTTTCGCTTCCG
CTTTCCAGCGTCGCGCAGGTGGCATCCTGATCGCTAGCCACCTGCAGAGCTTCCTGGAAGTTTCCTACCGTGTTTTCCGTCACCTG
GCTCAGCCGTGA

FIG._11B

Core4v
ATGACTCCATTAGGTCCAGCTTCCTCTCTGCCGCAAAGCTTCCTGCTGAAACTGCTGGAACAGATCCGTAAAATCCAGGGTGATGC
AGCTGCTCTGCAGGAAAAACTGTGCGCTACCTACAAACTGTGCCATCCGGAAGAACTGGTTCTGCTGGGTCACTCCCTGGGTATCC
CGTGGGCGCCGCTGAGCTCCTGCCCGAGCCAGGCTCTGCAGCTGGCTGGTTGCCTGTCCCAATTCCACAGCGGCCTTTTCCTGTTC
CAGGGTCTGCTGCAAGCTCTGGAAGGTATCTCCCCGGAACTGGGTCCGACCCTGGACACTCTGCAGCTGGACATCGCTGACCTGGC
TACCACCATCTGGCAGCAGATGGAAGAACTGGGTATGGCTCCGGCTCTGCAGCCGACCCAGGGTGCTATGCCGGCTTTCGCTTCCG
CTTTCCAGCGTCGCGCAGGTGGCATCCTGATCGCTAGCCACATCCAGAGCTGGTTCGAAGTTTCCTACCGTGCTTTCCGTCACCTG
GCTCAGCCGTGA

FIG._11C

FIG._12

-Melting Temperature (T_m)

	T_m (°C)	Extinction Coefficient (M^{-1} cm^{-1})
hG-CSFwt	60	15720
core4	72	14230
core4v	61	19730
core3	58	14230
sm0*	63	15720
fm4*	63	15720
fm7*	70	14230

* Derived by reverting some core4 or core3 mutant positions to wild type

FIG._16

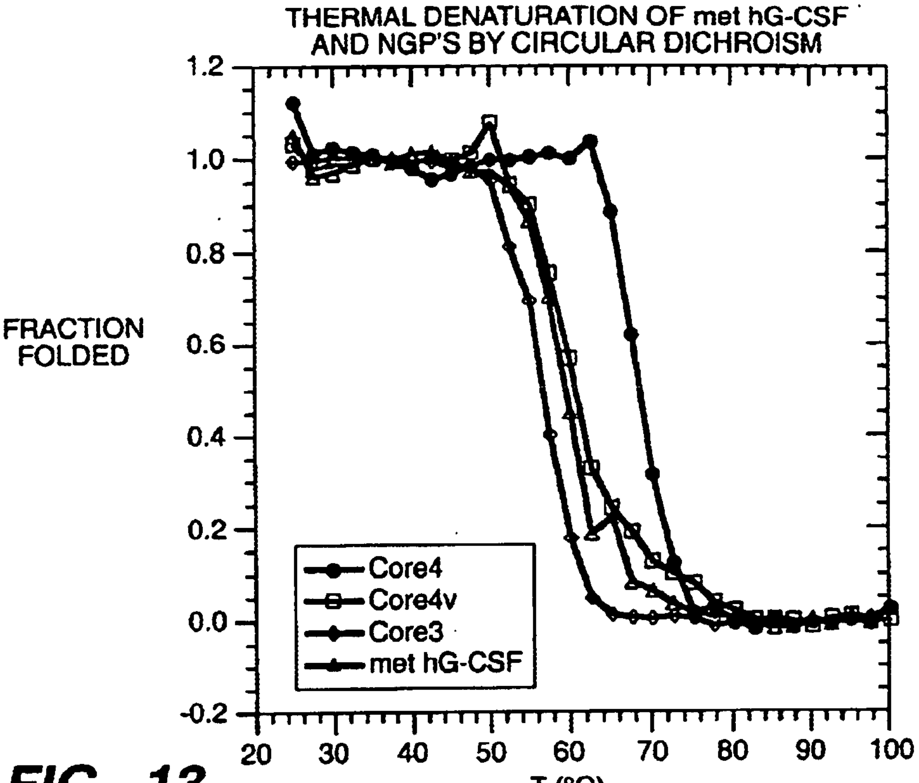

THERMAL DENATURATION OF met hG-CSF
AND NGP'S BY CIRCULAR DICHROISM
FRACTION FOLDED
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
20
30
40
50
60
70
80
90
100
T (ºC)
Core4
Core4v
Core3
met hG-CSF

FIG._13

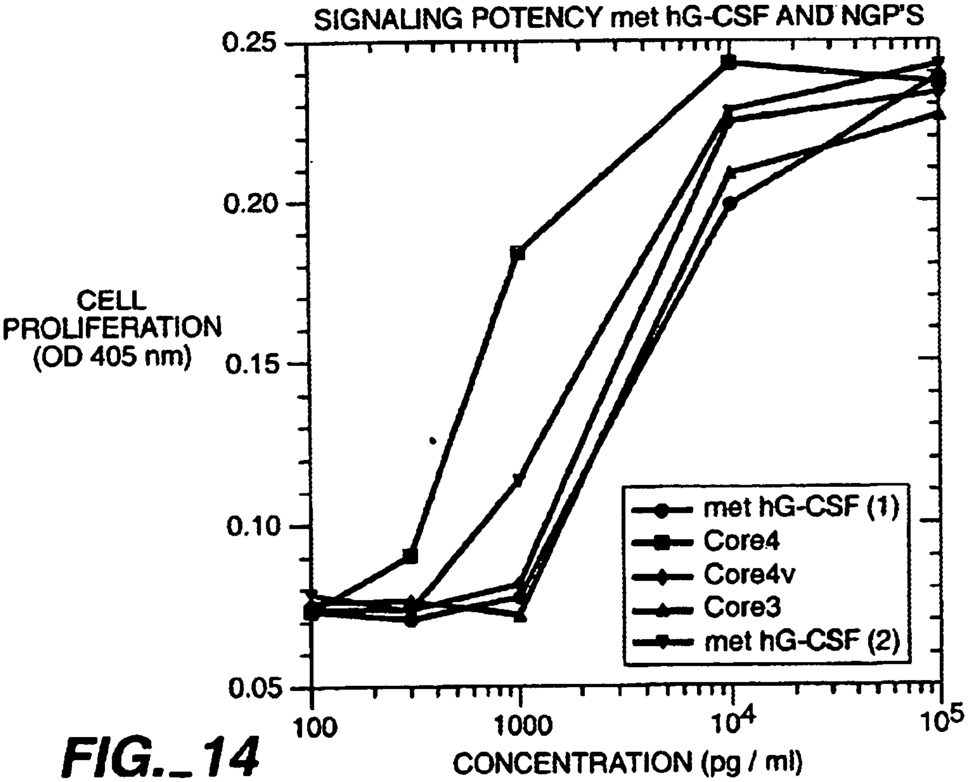

SIGNALING POTENCY met hG-CSF AND NGP'S
CELL PROLIFERATION (OD 405 nm)
0.25
0.20
0.15
0.10
0.05
100
1000
10^4
10^5
CONCENTRATION (pg / ml)
met hG-CSF (1)
Core4
Core4v
Core3
met hG-CSF (2)

FIG._14

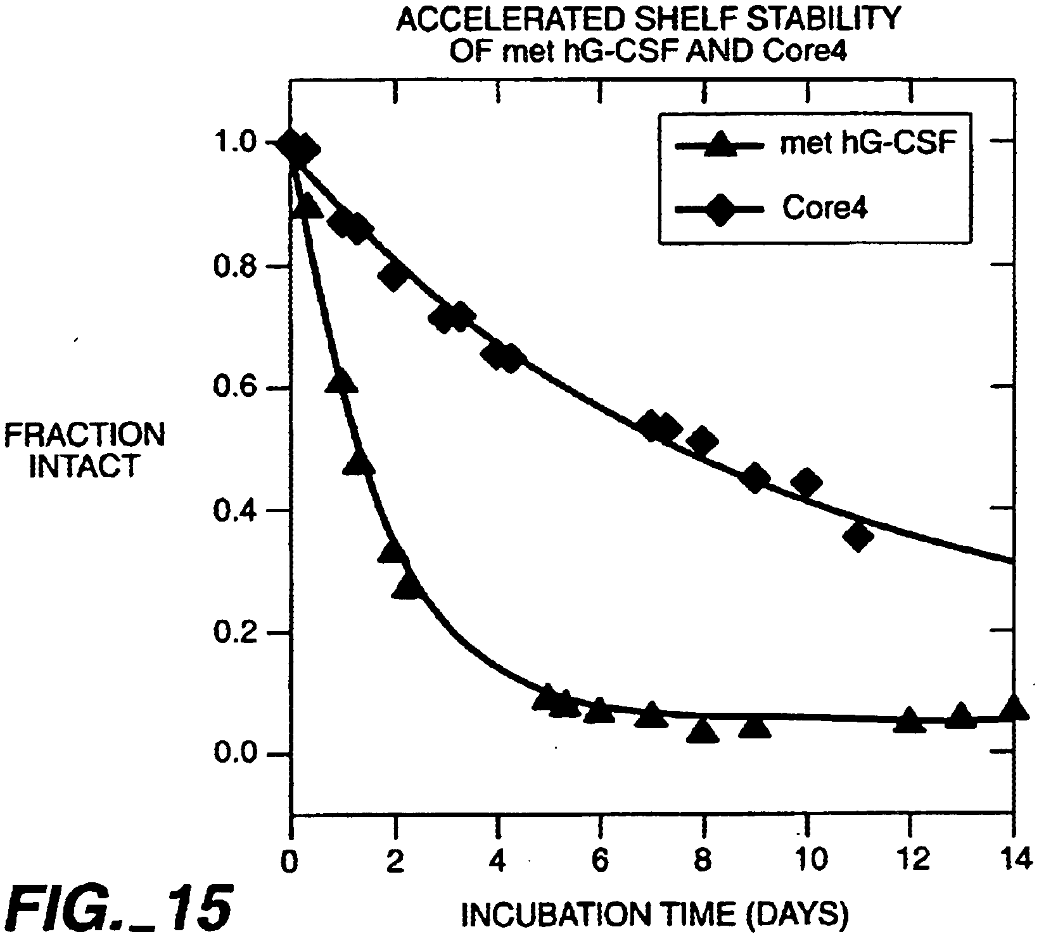

ACCELERATED SHELF STABILITY
OF met hG-CSF AND Core4
FRACTION INTACT
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
INCUBATION TIME (DAYS)
▲ met hG-CSF
◆ Core4

FIG._15

NUCLEIC ACIDS AND PROTEIN VARIANTS OF HG-CSF WITH GRANULOPOIETIC ACTIVITY

This application is a continuing application of U.S. Ser. Nos. 60/115,131, filed Jan. 6, 1999 and of 60/118,831, filed Feb. 5, 1999.

FIELD OF THE INVENTION

The invention relates to novel granulopoietic activity (GPA) proteins and nucleic acids. The invention further relates to the use of the GPA proteins in the treatment of G-CSF related disorders.

BACKGROUND OF THE INVENTION

The colony stimulating factors are a class of protein hormones that stimulate the proliferation and the function of specific blood cell types such as granulocytes. Granulocytes engulf and devour microbial invaders and cell debris and thus are crucial to infection response. Granulocytes have only a 6–12 hour life span in the bloodstream and are destroyed as they function. Accordingly, it necessary for the blood marrow stem cells to rapidly and constantly generate granulocytes. Granulocyte colony stimulating factor (G-CSF) is a protein that is essential for the proliferation and differentiation of granulocytes, particularly neutrophils.

However, as a result of their fast turnover, the granulocyte count falls rapidly and markedly upon bone marrow damage, for example from treatment with traditional cancer treatments, including chemotherapeutic agents and radiation, or immunologic disorders including AIDS. Accordingly, treatment with hG-CSF has been shown to be efficacious in minimizing some of the side effects of cancer therapies, as well as in treatment of suppressed immune systems.

However, wild-type hG-CSF has several disadvantages, including storage stability problems as well as a short half-life in the blood stream.

To this end, variants of G-CSF are known; see for example U.S. Pat. Nos. 5,214,132; 5,399,345; 5,790,421; 5,581,476; 4,999,291; 4,810,643; 4,833,127; 5,218,092; 5,362,853; 5,830,705; 5,580,755; 5,399,345 and 5,416,195 and references cited therein.

However, a need still exists for proteins exhibiting both significant stability and granulopoietic activity. Accordingly, it is an object of the invention to provide granulopoietic activity (GPA) proteins, nucleic acids and antibodies for the treatment of neutrophil disorders.

SUMMARY OF THE INVENTION

In accordance with the objects outlined above, the present invention provides non-naturally occurring GPA proteins (e.g. the proteins are not found in nature) comprising amino acid sequences that are less than about 95–97% identical to hG-CSF. The GPA proteins have at least one biological property of a G-CSF protein; for example, the GPA proteins will stimulate cells with a G-CSF receptor to proliferate. Thus the invention provides GPA proteins with amino acid sequences that have at least about 5 amino acid substitutions as compared to the hG-CSF sequence shown in FIG. 1.

In a further aspect, the present invention provides non-naturally occurring GPA conformers that have three dimensional backbone structures that substantially correspond to the three dimensional backbone structure of hG-CSF. The amino acid sequence of the conformer and the amino acid sequence of the hG-CSF are less than about 95% identical. In one aspect, at least about 90% of the non-identical amino acids are in a core region of the conformer. In other aspects, the conformer have at least about 100% of the non-identical amino acids are in a core region of the conformer.

In an additional aspect, the changes are selected from the amino acid residues at positions selected from 14, 17, 20, 21, 24, 27, 28, 31, 32, 34, 38, 78, 79, 85, 89, 91, 99, 102, 103, 107, 109, 110, 113, 116, 120, 145, 146, 147, 148, 151, 153, 155, 156, 157, 160, 161, 164, 168 and 170. Preferred embodiments include at least about 5 or 10 variations.

In a further aspect, the invention provides recombinant nucleic acids encoding the non-naturally occurring GPA proteins, expression vectors comprising the recombinant nucleic acids, and host cells comprising the recombinant nucleic acids and expression vectors.

In an additional aspect, the invention provides methods of producing the GPA proteins of the invention comprising culturing host cells comprising the recombinant nucleic acids under conditions suitable for expression of the nucleic acids. The proteins may optionally be recovered.

In a further aspect, the invention provides pharmaceutical compositions comprising a GPA protein of the invention and a pharmaceutical carrier.

In an additional aspect, the invention provides methods for treating a G-CSF responsive condition comprising administering a GPA protein of the invention to a patient. The C-CSF condition may be myelo-suppresive therapy, chronic neutropenia, or peripheral blood progenitor cell collection.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 depicts the nucleic acid (SEQ ID NO:1) and amino acid (SEQ ID NO:2) sequences of human G-CSF.

FIG. 2 depicts the variable residues in each GPA set.

FIG. 3 (SEQ ID NOS:2–15) depicts some preferred GPA sequences. The top line (SEQ ID NO:15) is the hG-CSF sequence . Any residue for which a change is not noted remains the same as the hG-CSF sequence. The second line (SEQ ID NO:3) is a GPA protein, bndry4_2, with variable boundary residues; 24 different positions were allowed to vary. The third line (SEQ ID NO:4) is a GPA protein, bndry4_core4, with boundary variable residues; this utilized 24 different boundary positions but used the optimal sequence from the core4 design as the starting template. The fourth line (SEQ ID NO:5) is a GPA protein, bndry4_AD, with boundary variable residues; however, the boundary residues were chosen on the outer two helices (A and D; 14 variable residue positions) since initial calculations suggested that the most pronounced changes in helical propensity result from modifications at these locations; improvements in helical propensity might lead to improved stability. The fifth line (SEQ ID NO:6) is a GPA protein, bndry4_AD_core4 with 14 variable boundary residues; again suing the optimal sequence from the core4 design as the starting template. The sixth line (SEQ ID NO:7) is a GPA protein, core4, that utilized 26 different variable core positions. The seventh line (SEQ ID NO:8) is a GPA protein, core4_V167A, that utilized 25 variable core positions. The eighth line (SEQ ID NO:9) is a GPA protein, core3, that had 34 core variable positions.

FIG. 4 depicts the Monte Carlo analysis of the core4 GPA sequence. At the left is shown the hG-CSF sequence position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns. At position 17, for example, the hG-CSF a mino acid is cysteine; in GPA proteins, 73.6% of the top 1000 sequences had leucine at this position, and 22.9% of the sequences had isoleucine.

FIG. 5 depicts the Monte Carlo analysis of the core4v GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns. At position 17, for example, the hG-CSF amino acid is cysteine; in GPA proteins, 69.7% of the top 1000 sequences had leucine at this position, and 5.1% of the sequences had valine; and 25.1% of the sequences had isoleucine.

FIG. 6 depicts the Monte Carlo analysis of the core3 GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns.

FIG. 7 depicts the Monte Carlo analysis of the bndry4_2 GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns.

FIG. 8 depicts the Monte Carlo analysis of the bndry4_core4 GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns.

FIG. 9 depicts the Monte Carlo analysis of the bndry4_AD GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns.

FIG. 10 depicts the Monte Carlo analysis of the bndry4_AD_core4 GPA sequence. At the left is shown the hG-CSF sequence; position numbers are shown in the second column, the ground state sequence is shown in the third column and the number of occurrences of all amino acids found in the top 1000 Monte Carlo sequences is shown in the last columns.

FIGS. 11A, 11B and 11C depict the gene sequences for three GPA proteins: FIG. 11A (SEQ ID NO:16) is the core3 GPA protein, FIG. 11B (SEQ ID NO:17) is the core4 GPA protein, and FIG. 11C (SEQ ID NO:18) is the core4v GPA protein.

FIG. 12 depicts the synthesis of a full-length gene and all possible mutations by PCR. Overlapping oligonucleotides corresponding to the full-length gene (black bar, Step 1) are synthesized, heated and annealed. Addition of Pfu DNA polymerase to the annealed oligonucleotides results in the 5' to 3' synthesis of DNA (Step 2) to produce longer DNA fragments (Step 3). Repeated cycles of heating, annealing (Step 4) results in the production of longer DNA, including some full-length molecules. These can be selected by a second round of PCR using primers (arrowed) corresponding to the end of the full-length gene (Step 5).

FIG. 13 depicts the thermal stability of met hG-CSF and several GPA proteins by circular dichroism (CD) spectroscopy. CD directly measures secondary structure content of a protein and can track the loss of structure in response to temperature or chemical denaturants. FIG. 13 shows the increased thermal stability of core4 relative to met hG-CSF.

FIG. 14 depicts the cell proliferation response to met hG-CSF and 3 novel GPA proteins. Cell proliferation of BaF/3 cells expressing hG-CSF receptor is shown as monitored by BrdU incorporation, plotted against protein concentration. BrdU incorporation is assessed by fluorescent ELISA. The figure shows the increased biological activity of core4 relative to met hG-CSF.

FIG. 15 depicts the kinetics of storage stability of met hG-CSF and core4 monitored by size exclusion chromatography HPLC. The two proteins were incubated in 5% sorbitol, 10 mM sodium acetate, 0.004% Tween-80 at pH 4.0 and and stored at 50° C. The protein concentration was 300 ug/ml. Monomeric protein was considered intact.

FIG. 16 depicts the melting temperature (Tm) and extinction coefficients of hG-CSF and some of the novel GPA proteins of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to novel proteins and nucleic acids possessing granulopoietic activity (sometimes referred to herein as "GPA proteins" and "GPA nucleic acids"). The proteins are generated using a system previously described in WO98/47089 and U.S. Ser. No. 09/127,926, both of which are expressly incorporated by reference in their entirety, that is a computational modeling system that allows the generation of extremely stable proteins without necessarily disturbing the biological functions of the protein itself. In this way, novel GPA proteins and nucleic acids are generated, that can have a plurality of mutations in comparison to the wild-type enzyme yet retain significant activity.

The computational method used to generate and evaluate the GPA proteins of the invention is briefly described as follows. In a preferred embodiment, the computational method used to generate the primary library is Protein Design Automaton (PDA), as is described in U.S. Ser. Nos. 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254, all of which are expressly incorporated herein by reference. Briefly, PDA can be described as follows. A known protein structure is used as the starting point. The residues to be optimized are then identified, which may be the entire sequence or subset(s) thereof. The side chains of any positions to be varied are then removed. The resulting structure consisting of the protein backbone and the remaining sidechains is called the template. Each variable residue position is then preferably classified as a core residue, a surface residue, or a boundary residue; each classification defines a subset of possible amino acid residues for the position (for example, core residues generally will be selected from the set of hydrophobic residues, surface residues generally will be selected from the hydrophilic residues, and boundary residues may be either). Each amino acid can be represented by a discrete set of all allowed conformers of each side chain, called rotamers. Thus, to arrive at an optimal sequence for a backbone, all possible sequences of rotamers must be screened, where each backbone position can be occupied either by each amino acid in all possible rotameric states, or a subset of amino acids, and thus a subset of rotamers.

Two sets of interactions are then calculated for each rotamer at every position: the interaction of the rotamer side chain with all or part of the backbone (the "singles" energy, also called the rotamer/template or rotamer/backbone energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position or a subset of the other positions (the "doubles" energy, also called the rotamer/rotamer energy). The energy of each of these interactions is calculated through the use of a variety of scoring functions, which include, but are not limited to, the energy of van der Waal's forces, the energy of hydrogen bonding, the energy of secondary structure propensity, the energy of surface area solvation and the electrostatics. Thus, the total energy of each rotamer interaction, both with the backbone and other rotamers, is calculated, and stored in a matrix form.

The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences, a number which grows exponentially with sequence length and renders the calculations either unwieldy or impossible in real time. Accordingly, to solve this combinatorial search problem, a "Dead End Elimination" (DEE) calculation is performed. The DEE calculation is based on the fact that if the worst total interaction of a first rotamer is still better than the best total interaction of a second rotamer, then the second rotamer cannot be part of the global optimum solution. Since the energies of all rotamers have already been calculated, the DEE approach only requires sums over the sequence length to test and eliminate rotamers, which speeds up the calculations considerably. DEE can be rerun comparing pairs of rotamers, or combinations of rotamers, which will eventually result in the determination of a single sequence which represents the global optimum energy.

Once the global solution has been found, a Monte Carlo search may be done to generate a rank-ordered list of sequences in the neighborhood of the DEE solution. Starting at the DEE solution, random positions are changed to other rotamers, and the new sequence energy is calculated. If the new sequence meets the criteria for acceptance, it is used as a starting point for another jump. After a predetermined number of jumps, a rank-ordered list of sequences is generated. In addition, as will be appreciated by those in the art, a Monte Carlo search may be done from a DEE run that is not completed; that is, a partial DEE run that has a number of sequences may be used to generate a Monte Carlo list.

As outlined in U.S. Ser. No. 09/127,926, the protein backbone (comprising (for a naturally occurring protein) the nitrogen, the carbonyl carbon, the α-carbon, and the carbonyl oxygen, along with the direction of the vector from the α-carbon to the β-carbon) may be altered prior to the computational analysis, by varying a set of parameters called supersecondary structure parameters.

Once a protein structure backbone is generated (with alterations, as outlined above) and input into the computer, explicit hydrogens are added if not included within the structure (for example, if the structure was generated by X-ray crystallography, hydrogens must be added). After hydrogen addition, energy minimizabon of the structure is run, to relax the hydrogens as well as the other atoms, bond angles and bond lengths. In a preferred embodiment, this is done by doing a number of steps of conjugate gradient minimizabon (Mayo et al., *J. Phys. Chem.* 94:8897 (1990)) of atomic coordinate positions to minimize the Dreiding force field with no electrostatics. Generally from about 10 to about 250 steps is preferred, with about 50 being most preferred.

The GPA backbone structure contains at least one variable residue position. Each GPA residue that can differ from the hG-CSF protein at an equivalent position is called a "variable residue". As is known in the art, the residues, or amino acids, of proteins are generally sequentially numbered starting with the N-terminus of the protein. Thus a protein having a methionine at it's N-terminus is said to have a methionine at residue or amino acid position 1, with the next residues as 2, 3, 4, etc. At each position, the wild type (i.e. naturally occurring) protein may have one of at least 20 amino acids, in any number of rotamers. By "variable residue position" herein is meant an amino acid position of the protein to be designed that is not fixed in the design method as a specific residue or rotamer, generally the wild-type hG-CSF residue or rotamer.

In a preferred embodiment, all of the residue positions of the protein are variable. That is, every amino acid side chain may be altered in the methods of the present invention.

In an alternate preferred embodiment, only some of the residue positions of the protein are variable, and the remainder are "fixed", that is, they are identified in the three dimensional structure as being a particular amino acid in a set conformation. In some embodiments, a fixed position is left in its original conformation (which may or may not correlate to a specific rotamer of the rotamer library being used). Alternatively, residues may be fixed as a non-wild type residue; for example, when known site-directed mutagenesis techniques have shown that a particular residue is desirable (for example, to eliminate a proteolytic site or alter the active site), the residue may be fixed as a particular amino acid. Alternatively, the methods of the present invention may be used to evaluate mutations de novo, as is discussed below. In an alternate preferred embodiment, a fixed position may be "floated"; the amino acid at that position is fixed, but different rotamers of that amino acid are tested. In this embodiment, the variable residues may be at least one, or anywhere from 0.1% to 99.9% of the total number of residues. Thus, for example, it may be possible to change only a few (or one) residues, or most of the residues, with all possibilities in between.

In a preferred embodiment, residues which can be fixed include, but are not limited to, structurally or biologically functional residues. For example, residues which are known to be important for biological activity, such as the residues which form the binding site for a binding partner (ligand/receptor, antigen/antibody, etc.), phosphorylation or glycosylation sites which are crucial to biological function, or structurally important residues, such as disulfide bridges, metal binding sites, critical hydrogen bonding residues, residues critical for backbone conformation such as proline or glycine, residues critical for packing interactions, etc. may all be fixed in a conformation or as a single rotamer, or "floated".

Similarly, residues which may be chosen as variable residues may be those that confer undesirable biological attributes, such as susceptibility to proteolytic degradation, dimerization or aggregation sites, glycosylation sites which may lead to immune responses, unwanted binding activity, unwanted allostery, undesirable biological activity but with a preservation of binding, etc.

In a preferred embodiment, each variable position is classified as either a core, surface or boundary residue position, although in some cases, as explained below, the variable position may be set to glycine to minimize backbone strain.

In one embodiment, only core residues are variable residues; alternate embodiments utilize methods for designing GPA proteins containing core, boundary and surface variable residues; core and surface variable residues; core and boundary variable residues; surface and boundary variable residues; as well as surface variable residues alone, or boundary variable residues alone. In general, preferred embodiments do not utilize surface variable residues, as this can lead to undesirable antigenicity; however, in applications that are not related to therapeutic use of the GPA proteins, it may be desirable to alter surface residues.

The classification of residue positions as core, surface or boundary may be done in several ways, as will be appreciated by those in the art and outlined in WO98/47089, hereby incorporated by reference in its entirety. In a preferred embodiment, the classification is done via a visual scan of the original protein backbone structure, including the side chains, and assigning a classification based on a subjective evaluation of one skilled in the art of protein modelling. Alternatively, a preferred embodiment utilizes an assessment of the orientation of the Cα-Cβ vectors relative to a solvent accessible surface computed using only the template Cα atoms. In a preferred embodiment, the solvent accessible surface for only the Cα atoms of the target fold is generated using the Connolly algorithm with a probe radius ranging from about 4 to about 12 Å, with from about 6 to about 10 Å being preferred, and 8 Å being particularly preferred. The Cα radius used ranges from about 1.6 Å to about 2.3 Å, with from about 1.8 to about 2.1 Å being preferred, and 1.95 Å being especially preferred. A residue is classified as a core position if a) the distance for its Cα, along its Cα-Cβ vector, to the solvent accessible surface is greater than about 4–6 Å, with greater than about 5.0 Å being especially preferred, and b) the distance for its Cβ to the nearest surface point is greater than about 1.5–3 Å, with greater than about 2.0 Å being especially preferred. The remaining residues are classified as surface positions if the sum of the distances from their Cα, along their Cα-Cβ vector, to the solvent accessible surface, plus the distance from their Cβ to the closest surface point was less than about 2.54 Å, with less than about 2.7 Å being especially preferred. All remaining residues are classified as boundary positions.

Suitable core and boundary positions for GPA proteins are outlined below.

Once each variable position is classified as either core, surface or boundary, a set of amino acid side chains, and thus a set of rotamers, is assigned to each position. That is, the set of possible amino acid side chains that the program will allow to be considered at any particular position is chosen. Subsequently, once the possible amino acid side chains are chosen, the set of rotamers that will be evaluated at a particular position can be determined. Thus, a core residue will generally be selected from the group of hydrophobic residues consisting of alanine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine (in some embodiments, when the α scaling factor of the van der Waals scoring function, described below, is low, methionine is removed from the set), and the rotamer set for each core position potentially includes rotamers for these eight amino acid side chains (all the rotamers if a backbone independent library is used, and subsets if a rotamer dependent backbone is used). Similarly, surface positions are generally selected from the group of hydrophilic residues consisting of alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine and histidine. The rotamer set for each surface position thus includes rotamers for these ten residues. Finally, boundary positions are generally chosen from alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine histidine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine. The rotamer set for each boundary position thus potentially includes every rotamer for these seventeen residues (assuming cysteine, glycine and proline are not used, although they can be). Additionally, in some preferred embodiments, a set of 18 naturally occurring amino acids (all except cysteine and proline, which are known to be particularly disruptive) are used.

Thus, as will be appreciated by those in the art, there is a computational benefit to classifying the residue positions, as it decreases the number of calculations. It should also be noted that there may be situations where the sets of core, boundary and surface residues are altered from those described above; for example, under some circumstances, one or more amino acids is either added or subtracted from the set of allowed amino acids. For example, some proteins which dimerize or multimerize, or have ligand binding sites, may contain hydrophobic surface residues, etc. In addition, residues that do not allow helix "capping" or the favorable interaction with an α-helix dipole may be subtracted from a set of allowed residues. This modification of amino acid groups is done on a residue by residue basis.

In a preferred embodiment, proline, cysteine and glycine are not included in the list of possible amino acid side chains, and thus the rotamers for these side chains are not used. However, in a preferred embodiment, when the variable residue position has a φ angle (that is, the dihedral angle defined by 1) the carbonyl carbon of the preceding amino acid; 2) the nitrogen atom of the current residue; 3) the α-carbon of the current residue; and 4) the carbonyl carbon of the current residue) greater than 0°, the position is set to glycine to minimize backbone strain.

Once the group of potential rotamers is assigned for each variable residue position, processing proceeds as outlined in U.S. Ser. No. 09/127,926 and PCT US98/07254. This processing step entails analyzing interactions of the rotamers with each other and with the protein backbone to generate optimized protein sequences. Simplistically, the processing initially comprises the use of a number of scoring functions to calculate energies of interactions of the rotamers, either to the backbone itself or other rotamers. Preferred PDA scoring functions include, but are not limited to, a Van der Waals potential scoring function, a hydrogen bond potential scoring function, an atomic solvation scoring function, a secondary structure propensity scoring function and an electrostatic scoring function. As is further described below, at least one scoring function is used to score each position, although the scoring functions may differ depending on the position classification or other considerations, like favorable interaction with an α-helix dipole. As outlined below, the total energy which is used in the calculations is the sum of the energy of each scoring function used at a particular position, as is generally shown in Equation 1:

$$E_{total}=nE_{vdw}+nE_{as}+nE_{h\text{-}bonding}+nE_{ss}+nE_{elec} \quad \text{Equation 1}$$

In Equation 1, the total energy is the sum of the energy of the van der Waals potential (E_{vdw}), the energy of atomic solvation (E_{as}), the energy of hydrogen bonding ($E_{h\text{-}bonding}$), the energy of secondary structure ($E_{ss}$and the energy of electrostatic interaction (E_{elec}). The term n is either 0 or 1, depending on whether the term is to be considered for the particular residue position.

As outlined in U.S. Ser. Nos. 60/061,097, 60/043,464, 60/054,678, 09/127,926 and PCT US98/07254, any combination of these scoring functions, either alone or in combination, may be used. Once the scoring functions to be used are identified for each variable position, the preferred first step in the computational analysis comprises the determination of the interaction of each possible rotamer with all or part of the remainder of the protein. That is, the energy of interaction, as measured by one or more of the scoring functions, of each possible rotamer at each variable residue position with either the backbone or other rotamers, is calculated. In a preferred embodiment, the interaction of each rotamer with the entire remainder of the protein, i.e. both the entire template and all other rotamers, is done. However, as outlined above, it is possible to only model a portion of a protein, for example a domain of a larger protein, and thus in some cases, not all of the protein need be considered.

In a preferred embodiment, the first step of the computational processing is done by calculating two sets of interactions for each rotamer at every position: the interaction of the rotamer side chain with the template or backbone (the "singles" energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position (the "doubles" energy), whether that position is varied or floated. It should be understood that the backbone in this case includes both the atoms of the protein structure backbone, as well as the atoms of any fixed residues, wherein the fixed residues are defined as a particular conformation of an amino acid.

Thus, "singles" (rotamer/template) energies are calculated for the interaction of every possible rotamer at every variable residue position with the backbone, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the rotamer and every hydrogen bonding atom of the backbone is evaluated, and the E_{HB} is calculated for each possible rotamer at every variable position. Similarly, for the van der Waals scoring function, every atom of the rotamer is compared to every atom of the template (generally excluding the backbone atoms of its own residue), and the E_{vdW} is calculated for each possible rotamer at every variable residue position. In addition, generally no van der Waals energy is calculated if the atoms are connected by three bonds or less. For the atomic solvation scoring function, the surface of the rotamer is measured against the surface of the template, and the E_{as} for each possible rotamer at every variable residue position is calculated. The secondary structure propensity scoring function is also considered as a singles energy, and thus the total singles energy may contain an E_{ss} term. As will be appreciated by those in the art, many of these energy terms will be close to zero, depending on the physical distance between the rotamer and the template position; that is, the farther apart the two moieties, the lower the energy.

For the calculation of "doubles" energy (rotamer/rotamer), the interaction energy of each possible rotamer is compared with every possible rotamer at all other variable residue positions. Thus, "doubles" energies are calculated for the interaction of every possible rotamer at every variable residue position with every possible rotamer at every other variable residue position, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the first rotamer and every hydrogen bonding atom of every possible second rotamer is evaluated, and the E_{HB} is calculated for each possible rotamer pair for any two variable positions. Similarly, for the van der Waals scoring function, every atom of the first rotamer is compared to every atom of every possible second rotamer, and the E_{vdW} is calculated for each possible rotamer pair at every two variable residue positions. For the atomic solvation scoring function, the surface of the first rotamer is measured against the surface of every possible second rotamer, and the E_{as} for each possible rotamer pair at every two variable residue positions is calculated. The secondary structure propensity scoring function need not be run as a "doubles" energy, as it is considered as a component of the "singles" energy. As will be appreciated by those in the art, many of these double energy terms will be close to zero, depending on the physical distance between the first rotamer and the second rotamer; that is, the farther apart the two moieties, the lower the energy.

Once the singles and doubles energies are calculated and stored, the next step of the computational processing may occur. As outlined in U.S. Ser. No. 09/127,926 and PCT US98/07254, preferred embodiments utilize a Dead End Elimination (DEE) step, and preferably a Monte Carlo step.

The computational processing results in a set of optimized GPA protein sequences. These optimized GPA protein sequences are generally significantly different from the wild-type hG-CSF sequence from which the backbone was taken.

Thus, in the broadest sense, the present invention is directed to GPA proteins that have granulopoietic activity. By "granulopoietic activity" or "GPA" herein is meant that the protein exhibits at least one, and preferably more, of the biological functions of a granulocyte-colony stimulating factor (G-CSF), as defined below.

By "protein" herein is meant at least two covalently attached amino acids, which includes proteins, polypeptides, oligopeptides and peptides. The protein may be made up of naturally occurring amino acids and peptide bonds, or synthetic peptidomimetic structures, generally depending on the method of synthesis. Thus "amino acid", or "peptide residue", as used herein means both naturally occurring and synthetic amino acids. For example, homo-phenylalanine, citrulline and noreleucine are considered amino acids for the purposes of the invention. "Amino acid" also includes imino acid residues such as proline and hydroxyproline. The side chains may be in either the (R) or the (S) configuration. In the preferred embodiment, the amino acids are in the (S) or L-configuration. If non-naturally occurring side chains are used, non-amino acid substituents may be used, for example to prevent or retard in vivo degradations. Proteins including non-naturally occurring amino acids may be synthesized or in some cases, made recombinantly; see van Hest et al., FEBS Lett 428:(1–2) 68–70 May 22, 1998 and Tang et al., Abstr. Pap Am. Chem. S218:U138-U138 Part 2 Aug. 22, 1999, both of which are expressly incorporated by reference herein.

The GPA proteins of the invention exhibit at least one biological function of a G-CSF. By "granulocyte colony stimulating factor" or "G-CSF" herein is meant a wild type G-CSF. The G-CSF may be from any number of organisms, with G-CSFs from mammals being particularly preferred. Suitable mammals include, but are not limited to, rodents (rats, mice, hamsters, guinea pigs, etc.), primates, farm animals (including sheep, goats, pigs, cows, horses, etc) and in the most preferred embodiment, from humans (this is sometimes referred to herein as hG-CSF, the sequence of which is depicted in FIG. 1). As will be appreciated by those in the art, GPAs based on G-CSFs from mammals other than humans may find use in animal models of human disease. The GI numbers for a variety of mammalian species is as follows: bovine 442671; dog 442673; sheep 310382; cat CAA69853; pig 2411469; mouse 309248; rat 1680659.

The GPA proteins of the invention exhibit at least one biological function of a G-CSF. By "biological function" or "biological property" herein is meant any one of the properties or functions of a G-CSF, including, but not limited to, the ability to stimulate cell proliferation, particularly of hematopoetic stem cells to produce granulocytes and particularly neutrophils; the ability to treat severe chronic neutropenia; the use in harvesting peripheral blood progenitor cells; the ability to enhance bone marrow transplantation therapy; as well as the stimulation of CFU -Gm type cells.

In a preferred embodiment, the biological function is granulopoietic activity (GPA). GPA is defined as the ability of the compound to stimulate cells that have a G-CSF receptor to proliferate. However, in some embodiments, GPA proteins may not possess GPA activity.

In a preferred embodiment, the assay system used to determine GPA is an in-vitro system as described in the examples, using Ba/F3 cells stably transfected with the gene encoding the human Class 1 G-CSF receptor; see Young et al. Protein Sci. 6:1228–1236 (1997), hereby expressly incorporated by reference in its entirety. In this system, cell proliferation is measured as a function of BrdU incorporation, which is incorporated into the nucleic acid of the proliferating cells. An increase above background of at least about 20%, with at least about 50% being preferred and at least about 100%, 500% and 1000% being especially preferred is an indication of GPA. An alternative assay is the CFU-GM cell assay as described in Zsebo et al, Immunobiology 172:175–184 (1986), also expressly incorporated by reference in its entirety.

In a preferred embodiment, an in-vivo system can be used to assay for GPA. For example, a suitable system is as described in U.S. Pat. No. 4,999,291, hereby incorporated by reference in its entirety. In general, in vivo assays require the administration of the GPA protein (or, in the case of gene therapy, of the GPA nucleic acid) to a suitable animal, followed by monitoring of the granulocyte count (or in some cases monitoring lymphocytes can be done) of the animal. In general, increases in neutrophil, granulocyte or lymphocyte counts without corresponding erythrocyte counts is indicative of G-CSF. Similarly, a useful in vivo assay system is as follows: male c57BL/6N mice are rendered neutropenic with a single intraperitoneal injection of 200 mg/kg cyclophosphamide (CPA). Beginning 24 hrs later and for 4 consecutive days from the day after the dosing with CPA, the mice are given a daily intravenous injection of 100 ug/kg of rhG-CSF, novel granulopoietic protein, or control vehicle. Granulopoietic activity is assayed on day 5 by bleeding the mice retro-orbitally and counting the number of white blood cells and polymorphonuclear neutrophils. See Hattori et al., Blood 75:1228–1233 (1990), expressly incorporated by reference in its entirety.

In a preferred embodiment, the antigenic profile in the host animal of the GPA protein is similar, and preferably identical, to the antigenic profile of the host G-CSF; that is, the GPA protein does not significantly stimulate the host organism (e.g. the patient) to an immune response; that is, any immune response is not clinically relevant and there is no allergic response or neutralization of the protein by an antibody. That is, in a preferred embodiment, the GPA protein does not contain additional or different epitopes from the G-CSF. By "epitope" or "determinant" herein is meant a portion of a protein which will generate and/or bind an antibody. Thus, in most instances, no significant amount of antibodies are generated to a GPA protein. In general, this is accomplished by not significantly altering surface residues, as outlined below nor by adding any amino acid residues on the surface which can become glycosylated, as novel glycosylation can result in an immune response.

The GPA proteins and nucleic acids of the invention are distinguishable from naturally occurring G-CSFs. A "naturally occurring G-CSF" is one that exists in nature and includes allelic variations; a representative sequence is the human sequence (hG-CSF) shown in FIG. 1. It should be noted that unless otherwise stated, all positional numbering is based on this human G-CSF sequence. That is, as will be appreciated by those in the art, an alignment of G-CSF proteins and GPA proteins can be done using standard programs, as is outlined below, with the identification of "equivalent" positions between the two proteins. Thus, the GPA proteins and nucleic acids of the invention are non-naturally occurring; that is, they do not exist in nature.

Thus, in a preferred embodiment, the GPA protein has an amino acid sequence that differs from a wild-type G-CSF sequence by at least 3% of the residues. That is, the GPA proteins of the invention are less than about 97% identical to a G-CSF amino acid sequence. Accordingly, a protein is a "GPA protein" if the overall homology of the protein sequence to the amino acid sequence shown in FIG. 1 is preferably less than about 97%, more preferably less than about 95%, even more preferably less than about 90% and most preferably less than 85%. In some embodiments the homology will be as low as about 75 to 80%. Stated differently, based on the hG-CSF sequence of 174 residues, GPA proteins have at least about 5 residues that differ from the hG-CSF sequence (3%), with GPA proteins having from 5 residues to upwards of 30 residues being different from the hG-CSF sequence. In some instances, GPA proteins have 3 or 4 different residues from the hG-CSF sequence. Preferred GPA proteins have 10–24 different residues with from about 10 to about 14 being particularly preferred (that is, 6–8% of the protein is not identical to hG-CSF).

Homology in this context means sequence similarity or identity, with identity being preferred. As is known in the art, a number of different programs can be used to identify whether a protein (or nucleic acid as discussed below) has sequence identity or similarity to a known sequence. Sequence identity and/or similarity is determined using standard techniques known in the art, including, but not limited to, the local sequence identity algorithm of Smith & Waterman, *Adv. Appl. Math.*, 2:482 (1981), by the sequence identity alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.*, 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Natl. Acad. Sci. U.S.A.*, 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Drive, Madison, Wis.), the Best Fit sequence program described by Devereux et al., *Nucl. Acid Res.*, 12:387–395 (1984), preferably using the default settings, or by inspection. Preferably, percent identity is calculated by FastDB based upon the following parameters: mismatch penalty of 1; gap penalty of 1; gap size penalty of 0.33; and joining penalty of 30, "Current Methods in Sequence Comparison and Analysis," Macromolecule Sequencing and Synthesis, Selected Methods and Applications, pp 127–149 (1988), Alan R. Liss, Inc.

An example of a useful algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Doolittle, J. Mol. Evol. 35:351–360 (1987); the method is similar to that described by Higgins & Sharp CABIOS 5:151–153 (1989). Useful PILEUP parameters including a default gap weight of 3.00, a default gap length weight of 0.10, and weighted end gaps.

Another example of a useful algorithm is the BLAST algorithm, described in Altschul et al., *J. Mol. Biol.*, 215, 403–410, (1990) and Karlin et al., *Proc. Natl. Acad. Sci. U.S.A.*, 90:5873–5787 (1993). A particularly useful BLAST program is the WU-BLAST-2 program which was obtained from Altschul et al., *Methods in Enzymology*, 266:460–480 (1996); http://blast.wustl/edu/blast/README.html]. WU-BLAST-2 uses several search parameters, most of which are set to the default values. The adjustable parameters are set with the following values: overlap span=1, overlap fraction =0.125, word threshold (T)=11. The HSP S and HSP S2 parameters are dynamic values and are established by the program itself depending upon the composition of the particular sequence and composition of the particular database against which the sequence of interest is being searched; however, the values may be adjusted to increase sensitivity.

An additional useful algorithm is gapped BLAST as reported by Altschul et al., *Nucl. Acids Res.*, 25:3389–3402. Gapped BLAST uses BLOSUM-62 substitution scores; threshold T parameter set to 9; the two-hit method to trigger ungapped extensions; charges gap lengths of k a cost of 10+k; X_u set to 16, and X_g set to 40 for database search stage and to 67 for the output stage of the algorithms. Gapped alignments are triggered by a score corresponding to ~22 bits.

A % amino acid sequence identity value is determined by the number of matching identical residues divided by the total number of residues of the "longer" sequence in the aligned region. The "longer" sequence is the one having the most actual residues in the aligned region (gaps introduced by WU-Blast-2 to maximize the alignment score are ignored).

In a similar manner, "percent (%) nucleic acid sequence identity" with respect to the coding sequence of the polypeptides identified herein is defined as the percentage of nucleotide residues in a candidate sequence that are identical with the nucleotide residues in the coding sequence of the cell cycle protein. A preferred method utilizes the BLASTN module of WU-BLAST-2 set to the default parameters, with overlap span and overlap fraction set to 1 and 0.125, respectively.

The alignment may include the introduction of gaps in the sequences to be aligned. In addition, for sequences which contain either more or fewer amino acids than the protein encoded by the sequence of FIG. **1**, it is understood that in one embodiment, the percentage of sequence identity will be determined based on the number of identical amino acids in relation to the total number of amino acids. Thus, for example, sequence identity of sequences shorter than that shown in FIG. **1**, as discussed below, will be determined using the number of amino acids in the shorter sequence, in one embodiment. In percent identity calculations relative weight is not assigned to various manifestations of sequence variation, such as, insertions, deletions, substitutions, etc.

In one embodiment, only identities are scored positively (+1) and all forms of sequence variation including gaps are assigned a value of "0", which obviates the need for a weighted scale or parameters as described below for sequence similarity calculations. Percent sequence identity can be calculated, for example, by dividing the number of matching identical residues by the total number of residues of the "shorter" sequence in the aligned region and multiplying by 100. The "longer" sequence is the one having the most actual residues in the aligned region.

Thus, GPA proteins of the present invention may be shorter or longer than the amino acid sequence shown in FIG. 1. Thus, in a preferred embodiment, included within the definition of GPA proteins are portions or fragments of the sequences depicted herein. Fragments of GPA proteins are considered GPA proteins if a) they share at least one antigenic epitope; b) have at least the indicated homology; c) and preferably have GPA biological activity as defined herein.

In a preferred embodiment, as is more fully outlined below, the GPA proteins include further amino acid variations, as compared to a wild-type G-CSF, than those outlined herein. In addition, as outlined herein, any of the variations depicted herein may be combined in any way to form additional novel GPA proteins.

In addition, GPA proteins can be made that are longer than those depicted in the figures, for example, by the addition of epitope or purification tags, as outlined herein, the addition of other fusion sequences, etc. For example, the GPA proteins of the invention may be fused to other therapeutic proteins such as IL-11 or to other proteins such as Fc or serum albumin for pharmacokinetic purposes. See for example U.S. Pat. Nos. 5,766,883 and 5,876,969, both of which are expressly incorporated by reference.

In a preferred embodiment, the GPA proteins comprise variable residues in core and boundary residues.

hG-CSF core residues are as follows: positions 17, 21, 24, 28, 31, 35, 41, 47, 54, 56, 75, 78, 82, 85, 88, 89, 92, 95, 99, 103, 106, 110, 113, 114, 117, 140, 149, 150, 151, 152, 153, 154, 157, 160, 161 and 168. Accordingly, in a preferred embodiment, GPA proteins have variable positions selected from these positions.

In a preferred embodiment, GPA proteins have variable positions selected solely from core residues of hG-CSF. Alternatively, at least a majority (51%) of the variable positions are selected from core residues, with at least about 75% of the variable positions being preferably selected from core residue positions, and at least about 90% of the variable positions being particularly preferred. A specifically preferred embodiment has only core variable positions altered as compared to hG-CSF.

Particularly preferred embodiments where GPA proteins have variable core positions as compared to hG-CSF are shown in the Figures.

In one embodiment, the variable core positions are altered to any of the other 19 amino acids. In a preferred embodiment, the variable core residues are chosen from Ala, Val, Phe, Ile, Leu, Tyr and Trp. hG-CSF boundary residues are as follows: positions 14, 20, 27, 32, 34, 38, 77, 79, 84, 91, 99, 102, 107, 109, 116, 120, 145, 146, 147, 155, 156, 164 and 170. Accordingly, in a preferred embodiment, GPA proteins have variable positions selected from these positions.

In a preferred embodiment, the boundary core positions are altered to any of the other 19 amino acids. In a preferred embodiment, the variable boundary residues are chose from Ala, Val, Leu, Ile, Asp, Asn, Glu, Gln, Lys, Ser, Thr and His (preferably protonated His).

In a preferred embodiment, the GPA protein of the invention has a sequence that differs from a wild-type G-CSF protein in at least one amino acid position selected from position 14, 17, 20, 21, 24, 27, 28, 31, 32, 34, 38, 78, 79, 85, 89, 91, 99, 102, 103, 107, 109, 110, 113, 116, 120, 145, 146, 147, 148, 151, 153, 155, 156, 157, 160, 161, 164, 168 and 170; see also FIG. **2** which outlines sets of amino acid positions.

Preferred amino acids for each position, including the hG-CSF residue, are shown in FIGS. **3–10**. Thus, for example, at position 17, preferred amino acids are Leu, Val and Ile; at position 21, Val, Ile, Phe, Ala, and Tyr; etc.

Preferred changes are as follows: Leu14Ile; Cys17Ala; Cys17Leu; Cys17Ile; Gln20Leu; Val2Ile; Val21Ala; Val21Phe; Val21Tyr; Ile24Ala; Ile24Val; Ile24Leu; Asp27Glu; Asp27Ser; Gly28Ala; Gly28Leu; Leu31Val; Gln32Leu; Gln32Val; Gln32Ile; Lys34Glu; Lys34Gln; Lys35Ile; Lys35Val; Thr38His; Thr38Val; Thr38Ile; Thr38Glu; Thr38Lys; Leu78Phe; Leu78Ala; Leu78Val; Leu78Ile; Leu78Tyr; His79Leu; Leu82Ala; Leu82Phe; Tyr85Val; Tyr85Ile; Tyr85Phe; Tyr85Trp; Leu89Phe; Leu89Trp; Ala91Lys; Leu92Phe; Leu99Glu; Thr102Lys; Thr102Val; Thr102Leu; Thr102Ile Thr102Glu; Thr102Gln; Leu103Val; Leu103Ile; Leu103Ala; Leu106Val; Gln107Ile; Gln107Val; Gln107Leu; Val109Glu; Val109Asp; Val109Gln; Val110Ala; Val110Leu; Val110Ile; Phe113Ala; Phe113Leu; Thr116Ile; Thr116Val; Thr116Leu; Thr116Glu; Thr116Ala; Ile117Val; Ile117Leu; Ile117Phe; Ile117Trp; Gln120Leu; Gln145Glu; Arg146Lys; Arg146Gln; Arg147Glu; Arg147Lys; Ala148Asp; Ala148Thr; Val151Ile; Val153Ile; Ser155Ile; His156Leu; Leu157Ala; Leu157Val; Leu157Ile; Phe160Trp; Leu161Phe; Ser164Ala; Leu157Ile; Phe160Trp; Leu161Phe; Leu161Ala; Leu161Val; Val167Ala; Leu168Phe; His170Asp; His170Leu; His170Glu; His170Gln; and His170Lys. These may be done either individually or in combination, with any combination being possible. However, as outlined herein, preferred embodiments utilize at least four, and preferably more, variable positions in each GPA protein.

Particularly preferred sequences are selected from the group consisting of: C17L, G28A, L78F, Y85F, L103V, V110I, F113L, V151I, V153I and L168F, SEQ ID NO: 7; and L14I, Q20L, D27E , Q32L, K34E, T38H, H79L, A91K, T102K, Q107I, D109E, T116I, Q120L, R146K, R147E, A148D, S155I, H156L, S163A, SEQ ID NO: 18.

In a preferred embodiment, the GPA proteins do not have sole single variable positions at positions 17, 24, 35, 41, 18, 68, 26, 174, 170, 167, 44, 47, 23, 20, 28, 127, 138, 13, 121 or 124. Similarly, preferred embodiments of GPA proteins do not only have two variable positions at 127 and 138 or 37 and 43. In a preferred embodiment, the GPA proteins do not have only three variable positions at 17, 24 and 41; 17, 24 and 35; and 17, 35 and 41. Furthermore, preferred GPA proteins doe not have only four variable positions at 17, 24, 35 and 41.

In a preferred embodiment, the GPA proteins of the invention are hG-CSF conformers. By "conformer" herein is meant a protein that has a protein backbone 3D structure that is virtually the same but has significant differences in the amino acid side chains. That is, the GPA proteins of the invention define a conformer set, wherein all of the proteins of the set share a backbone structure and yet have sequences that differ by at least 3–5%. "Backbone" in this context means the non-side chain atoms: the nitrogen, carbonyl carbon and oxygen, and the α-carbon, and the hydrogens attached to the nitrogen and α-carbon. To be considered a conformer, a protein must have backbone atoms that are no more than 2 Å from the hG-CSF structure, with no more than 1.5 Å being preferred, and no more than 1 Å being particularly preferred. In general, these distances may be determined in two ways. In one embodiment, each potential conformer is crystallized and its three dimensional structure determined. Alternatively, as the former is quite tedious, the sequence of each potential conformer is run in the PDA program to determine whether it is a conformer.

GPA proteins may also be identified as being encoded by GPA nucleic acids. In the case of the nucleic acid, the overall homology of the nucleic acid sequence is commensurate with amino acid homology but takes into account the degeneracy in the genetic code and codon bias of different organisms. Accordingly, the nucleic acid sequence homology may be either lower or higher than that of the protein sequence, with lower homology being preferred.

In a preferred embodiment, an GPA nucleic acid encodes an GPA protein. As will be appreciated by those in the art, due to the degeneracy of the genetic code, an extremely large number of nucleic acids may be made, all of which encode the GPA proteins of the present invention. Thus, having identified a particular amino acid sequence, those skilled in the art could make any number of different nucleic acids, by simply modifying the sequence of one or more codons in a way which does not change the amino acid sequence of the GPA.

In one embodiment, the nucleic acid homology is determined through hybridization studies. Thus, for example, nucleic acids which hybridize under high stringency to the nucleic acid sequences shown in FIG. 1 or its complement and encode a GPA protein is considered an GPA gene.

High stringency conditions are known in the art; see for example Maniatis et al., Molecular Cloning: A Laboratory Manual, 2d Edition, 1989, and Short Protocols in Molecular Biology, ed. Ausubel, et al., both of which are hereby incorporated by reference. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Techniques in Biochemistry and Molecular Biology—Hybridization with Nucleic Acid Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, stringent conditions are selected to be about 5–10° C. lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength, pH and nucleic acid concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T_m, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30° C. for short probes (e.g. 10 to 50 nucleotides) and at least about 60° C. for long probes (e.g. greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

In another embodiment, less stringent hybridization conditions are used; for example, moderate or low stringency conditions may be used, as are known in the art; see Maniatis and Ausubel, supra, and Tijssen, supra.

The GPA proteins and nucleic acids of the present invention are recombinant. As used herein, "nucleic acid" may refer to either DNA or RNA, or molecules which contain both deoxy- and ribonucleotides. The nucleic acids include genomic DNA, cDNA and oligonucleotides including sense and anti-sense nucleic acids. Such nucleic acids may also contain modifications in the ribose-phosphate backbone to increase stability and half life of such molecules in physiological environments.

The nucleic acid may be double stranded, single stranded, or contain portions of both double stranded or single stranded sequence. As will be appreciated by those in the art, the depiction of a single strand ("Watson") also defines the sequence of the other strand ("Crick"); thus the sequence depicted in FIG. 1 also includes the complement of the sequence. By the term "recombinant nucleic acid"herein is meant nucleic acid, originally formed in vitro, in general, by the manipulation of nucleic acid by endonucleases, in a form not normally found in nature. Thus an isolated GPA nucleic acid, in a linear form, or an expression vector formed in vitro by ligating DNA molecules that are not normally joined, are both considered recombinant for the purposes of this invention. It is understood that once a recombinant nucleic acid is made and reintroduced into a host cell or organism, it will replicate non-recombinantly, i.e. using the in vivo cellular machinery of the host cell rather than in vitro manipulations; however, such nucleic acids, once produced recombinantly, although subsequently replicated non-recombinantly, are still considered recombinant for the purposes of the invention.

Similarly, a "recombinant protein" is a protein made using recombinant techniques, i.e. through the expression of a recombinant nucleic acid as depicted above. A recombinant protein is distinguished from naturally occurring protein by at least one or more characteristics. For example, the protein may be isolated or purified away from some or all of the proteins and compounds with which it is normally associated in its wild type host, and thus may be substantially pure. For example, an isolated protein is unaccompanied by at least some of the material with which it is normally associated in its natural state, preferably constituting at least about 0.5%, more preferably at least about 5% by weight of the total protein in a given sample. A substantially pure protein comprises at least about 75% by weight of the total protein, with at least about 80% being preferred, and at least about 90% being particularly preferred. The definition includes the production of an GPA protein from one organism in a different organism or host cell. Alternatively, the protein may be made at a significantly higher concentration than is normally seen, through the use of a inducible promoter or high expression promoter, such that the protein is made at increased concentration levels. Furthermore, all of the GPA proteins outlined herein are in a form not normally found in nature, as they contain amino acid substitutions, insertions and deletions, with substitutions being preferred, as discussed below.

Also included within the definition of GPA proteins of the present invention are amino acid sequence variants of the GPA sequences outlined herein and shown in the Figures. That is, the GPA proteins may contain additional variable positions as compared to hG-CSF. These variants fall into one or more of three classes: substitutional, insertional or deletional variants. These variants ordinarily are prepared by site specific mutagenesis of nucleotides in the DNA encoding a GPA protein, using cassette or PCR mutagenesis or other techniques well known in the art, to produce DNA encoding the variant, and thereafter expressing the DNA in recombinant cell culture as outlined above. However, variant GPA protein fragments having up to about 100–150 residues may be prepared by in vitro synthesis using established techniques. Amino acid sequence variants are characterized by the predetermined nature of the variation, a feature that sets them apart from naturally occurring allelic or interspecies variation of the GPA protein amino acid sequence. The variants typically exhibit the same qualitative biological activity as the naturally occurring analogue, although variants can also be selected which have modified characteristics as will be more fully outlined below.

While the site or region for introducing an amino acid sequence variation is predetermined, the mutation per se need not be predetermined. For example, in order to optimize the performance of a mutation at a given site, random mutagenesis may be conducted at the target codon or region and the expressed GPA variants screened for the optimal combination of desired activity. Techniques for making substitution mutations at predetermined sites in DNA having a known sequence are well known, for example, M13 primer mutagenesis and PCR mutagenesis. Screening of the mutants is done using assays of GPA protein activities.

Amino acid substitutions are typically of single residues; insertions usually will be on the order of from about 1 to 20 amino acids, although considerably larger insertions may be tolerated. Deletions range from about 1 to about 20 residues, although in some cases deletions may be much larger.

Substitutions, deletions, insertions or any combination thereof may be used to arrive at a final derivative. Generally these changes are done on a few amino acids to minimize the alteration of the molecule. However, larger changes may be tolerated in certain circumstances. When small alterations in the characteristics of the GPA protein are desired, substitutions are generally made in accordance with the following chart:

Chart 1

Original Residue	Exemplary Substitutions
Ala	Ser
Arg	Lys
Asn	Gln, His
Asp	Glu
Cys	Ser, Ala
Gln	Asn
Glu	Asp
Gly	Pro
His	Asn, Gln
Ile	Leu, Val
Leu	Ile, Val
Lys	Arg, Gln, Glu
Met	Leu, Ile
Phe	Met, Leu, Tyr
Ser	Thr
Thr	Ser
Trp	Tyr
Tyr	Trp, Phe
Val	Ile, Leu

Substantial changes in function or immunological identity are made by selecting substitutions that are less conservative than those shown in Chart I. For example, substitutions may be made which more significantly affect: the structure of the polypeptide backbone in the area of the alteration, for example the alpha-helical or beta-sheet structure; the charge or hydrophobicity of the molecule at the target site; or the bulk of the side chain. The substitutions which in general are expected to produce the greatest changes in the polypeptide's properties are those in which (a) a hydrophilic residue, e.g. seryl or threonyl, is substituted for (or by) a hydrophobic residue, e.g. leucyl, isoleucyl, phenylalanyl, valyl or alanyl; (b) a cysteine or proline is substituted for (or by) any other residue; (c) a residue having an electropositive side chain, e.g. lysyl, arginyl, or histidyl, is substituted for (or by) an electronegative residue, e.g. glutamyl or aspartyl; or (d) a residue having a bulky side chain, e.g. phenylalanine, is substituted for (or by) one not having a side chain, e.g. glycine.

The variants typically exhibit the same qualitative biological activity and will elicit the same immune response as the original GPA protein, although variants also are selected to modify the characteristics of the GPA proteins as needed. Alternatively, the variant may be designed such that the biological activity of the GPA protein is altered. For example, glycosylation sites may be altered or removed.

Similarly, the biological function may be altered; for example, in some instances it may be desirable to have more or less potent granulopoietic activity.

The GPA proteins and nucleic acids of the invention can be made in a number of ways. As will be appreciated by those in the art, it is possible to synthesize proteins using standard techniques well known in the art. See for example Wilken et al., Curr. Opin. Biotechnol. 9:412–26 (1998), hereby expressly incorporated by reference.

Alternatively, and preferably, the proteins and nucleic acids of the invention are made using recombinant techniques. Using the nucleic acids of the present invention which encode a GPA protein, a variety of expression vectors are made. The expression vectors may be either self-replicating extrachromosomal vectors or vectors which integrate into a host genome. Generally, these expression vectors include transcriptional and translational regulatory nucleic acid operably linked to the nucleic acid encoding the GPA protein. The term "control sequences" refers to DNA sequences necessary for the expression of an operably linked coding sequence in a particular host organism.

The control sequences that are suitable for prokaryotes, for example, include a promoter, optionally an operator sequence, and a ribosome binding site. Eukaryotc cells are known to utilize promoters, polyadenylation signals, and enhancers.

Nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For example, DNA for a presequence or secretory leader is operably linked to DNA for a polypeptide if it is expressed as a preprotein that participates in the secretion of the polypeptide; a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the sequence; or a ribosome binding site is operably linked to a coding sequence if it is positioned so as to facilitate translation. Generally, "operably linked" means that the DNA sequences being linked are contiguous, and, in the case of a secretory leader, contiguous and in reading phase. However, enhancers do not have to be contiguous. Linking is accomplished by ligation at convenient restriction sites. If such sites do not exist, the synthetic oligonucleotide adaptors or linkers are used in accordance with conventional practice. The transcriptional and translational regulatory nucleic acid will generally be appropriate to the host cell used to express the fusion protein; for example, transcriptional and translational regulatory nucleic acid sequences from Bacillus are preferably used to express the fusion protein in Bacillus. Numerous types of appropriate expression vectors, and suitable regulatory sequences are known in the art for a variety of host cells.

In general, the transcriptional and translational regulatory sequences may include, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, and enhancer or activator sequences. In a preferred embodiment, the regulatory sequences include a promoter and transcriptional start and stop sequences.

Promoter sequences encode either constitutive or inducible promoters. The promoters may be either naturally occurring promoters or hybrid promoters. Hybrid promoters, which combine elements of more than one promoter, are also known in the art, and are useful in the present invention. In a preferred embodiment, the promoters are strong promoters, allowing high expression in cells, particularly mammalian cells, such as the CMV promoter, particularly in combination with a Tet regulatory element.

In addition, the expression vector may comprise additional elements. For example, the expression vector may have two replication systems, thus allowing it to be maintained in two organisms, for example in mammalian or insect cells for expression and in a procaryotic host for cloning and amplification. Furthermore, for integrating expression vectors, the expression vector contains at least one sequence homologous to the host cell genome, and preferably two homologous sequences which flank the expression construct. The integrating vector may be directed to a specific locus in the host cell by selecting the appropriate homologous sequence for inclusion in the vector. Constructs for integrating vectors are well known in the art.

In addition, in a preferred embodiment, the expression vector contains a selectable marker gene to allow the selection of transformed host cells. Selection genes are well known in the art and will vary with the host cell used.

A preferred expression vector system is a retroviral vector system such as is generally described in PCT/US97/01019 and PCT/US97/01048, both of which are hereby expressly incorporated by reference.

The GPA nucleic acids are introduced into the cells. By "introduced into " or grammatical equivalents herein is meant that the nucleic acids enter the cells in a manner suitable for subsequent expression of the nucleic acid. The method of introduction is largely dictated by the targeted cell type, discussed below. Exemplary methods include $CaPO_4$ precipitation, liposome fusion, lipofectin®, electroporation, viral infection, etc. The GPA nucleic acids may stably integrate into the genome of the host cell (for example, with retroviral introduction, outlined below), or may exist either transiently or stably in the cytoplasm (i.e. through the use of traditional plasmids, utilizing standard regulatory sequences, selection markers, etc.).

The GPA proteins of the present invention are produced by culturing a host cell transformed with an expression vector containing nucleic acid encoding a GPA protein, under the appropriate conditions to induce or cause expression of the GPA protein. The conditions appropriate for GPA protein expression will vary with the choice of the expression vector and the host cell, and will be easily ascertained by one skilled in the art through routine experimentation. For example, the use of constitutive promoters in the expression vector will require optimizing the growth and proliferation of the host cell, while the use of an inducible promoter requires the appropriate growth conditions for induction. In addition, in some embodiments, the timing of the harvest is important. For example, the baculoviral systems used in insect cell expression are lytic viruses, and thus harvest time selection can be crucial for product yield.

Appropriate host cells include yeast, bacteria, archebacteria, fungi, and insect and animal cells, including mammalian cells. Of particular interest are *Drosophila melangaster* cells, *Saccharomyces cerevisiae* and other yeasts, *E. coli*, *Bacillus subtilis*, SF9 cells, C129 cells, 293 cells, Neurospora, BHK, CHO, COS, *Pichia Pastoris*, etc.

In a preferred embodiment, the GPA proteins are expressed in mammalian cells. Mammalian expression systems are also known in the art, and include retroviral systems. A mammalian promoter is any DNA sequence capable of binding mammalian RNA polymerase and initiating the downstream (3') transcription of a coding sequence for the fusion protein into mRNA. A promoter will have a transcription initiating region, which is usually placed proximal to the 5' end of the coding sequence, and a TATA box, using a located 25–30 base pairs upstream of the transcription initiation site. The TATA box is thought to direct RNA polymerase II to begin RNA synthesis at the correct site. A mammalian promoter will also contain an upstream promoter element (enhancer element), typically located within 100 to 200 base pairs upstream of the TATA box. An upstream promoter element determines the rate at which transcription is initiated and can act in either orientation. Of particular use as mammalian promoters are the promoters from mammalian viral genes, since the viral genes are often highly expressed and have a broad host range. Examples include the SV40 early promoter, mouse mammary tumor virus LTR promoter, adenovirus major late promoter, herpes simplex virus promoter, and the CMV promoter.

Typically, transcription termination and polyadenylation sequences recognized by mammalian cells are regulatory regions located 3' to the translation stop codon and thus, together with the promoter elements, flank the coding sequence. The 3' terminus of the mature mRNA is formed by site-specific post-translational cleavage and polyadenylation. Examples of transcription terminator and polyadenylation signals include those derived form SV40.

In a preferred embodiment, when combinations of variable positions are to be made, the nucleic acids encoding the GPA proteins are made using a variety of combinatorial techniques. For example, "shuffling" techniques such as are outlined in U.S. Pat. Nos. 5,811,238; 5,605,721 and 5,830,721, and related patents, all of which are hereby expressly incorporated by reference.

In a preferred embodiment, multiple PCR reactions with pooled oligonucleotides is done, as is generally depicted in FIG. 12. In this embodiment, overlapping oligonucleotides are synthesized which correspond to the full length gene. Again, these oligonucleotides may represent all of the different amino acids at each variant position or subsets.

In a preferred embodiment, these oligonucleotides are pooled in equal proportions and multiple PCR reactions are performed to create full length sequences containing the combinations of variable positions.

In a preferred embodiment, the different oligonucleotides are added in relative amounts corresponding to a probability distribution table; that is, as shown in FIGS. 3–10, different amino acids have different probabalistic chances of being at a particular position. Thus, for example, as shown in FIG. 4, out of the top 1000 sequences, position 103 has valine 35% of the time, leucine 26% of the time, and isoleucine 31% of the time. The multiple PCR reactions thus result in full length sequences with the desired combinations of variable amino acids in the desired proportions.

The total number of oligonucleotides needed is a function of the number of positions being mutated and the number of mutations being considered at these positions:

(number of oligos for constant positions)+M1+M2+M3+. . .
Mn=(total number of oligos required)

where Mn is the number of amino acids considered at position n in the sequence.

In a preferred embodiment, each overlapping oligonucleotide comprises only one position to be varied; in alternate embodiments, the variant positions are too close together to allow this and multiple variants per oligonucleotide are used to allow complete recombination of all the possibilities. That is, each oligo can contain the codon for a single position being varied, or for more than one position being varied. The multiple positions being varied must be close in sequence to prevent the oligo length from being impractical. For multiple variable positions on an oligonucleotide, particular combinations of variable residues can be included or excluded in the library by including or excluding the oligonucleotide encoding that combination. The total number of oligonucleotides required increases when multiple variable positions are encoded by a single oligonucleotide. The annealed regions are the ones that remain constant, i.e. have the sequence of the reference sequence.

Oligonucleotides with insertions or deletions of codons can be used to create a library expressing different length proteins. In particular computational sequence screening for insertions or deletions can result in secondary libraries defining different length proteins, which can be expressed by a library of pooled oligonucleotide of different lengths.

In a preferred embodiment, error-prone PCR is done. See U.S. Pat. Nos. 5,605,793, 5,811,238, and 5,830,721, all of which are hereby incorporated by reference. This can be done on the optimal sequence or on top members of the GPA set. In this embodiment, the gene for the optimal GPA sequence found in the computational screen can be synthesized. Error prone PCR is then performed on the optimal sequence gene in the presence of oligonucleotides that code for the variable residues at the variant positions (bias oligonucleotides). The addition of the oligonucleotides will create a bias favoring the incorporation of the variations in the secondary library. Alternatively, only oligonucleotides for certain variations may be used to bias the library.

In a preferred embodiment, error-prone PCR in combination with the overlapping oligonucleotide method outlined in FIG. 12 is done.

In a preferred embodiment, gene shuffling with error prone PCR can be performed on the gene for the optimal sequence, in the presence of bias oligonucleotides, to create a DNA sequence library that reflects the proportion of the variations. The choice of the bias oligonucleotides can be done in a variety of ways; they can chosen on the basis of their frequency, i.e. oligonucleotides encoding high variation frequency positions can be used; alternatively, oligonucleotides containing the most variable positions can be used, such that the diversity is increased; if the GPA protein set is ranked, some number of top scoring positions can be used to generate bias oligonucleotides; random positions may be chosen; a few top scoring and a few low scoring ones may be chosen; etc. What is important is to generate new sequences based on preferred variable positions and sequences. Similarly, a top set of GPA proteins may be "shuffled" using traditional shuffling methods or the overlapping oligonucleotide methods of FIG. 12.

The methods of introducing exogenous nucleic acid into mammalian hosts, as well as other hosts, is well known in the art, and will vary with the host cell used. Techniques include dextran-mediated transfection, calcium phosphate precipitation, polybrene mediated transfection, protoplast fusion, electroporation, viral infection, encapsulation of the polynucleotide(s) in liposomes, and direct microinjection of the DNA into nuclei. As outlined herein, a particularly preferred method utilizes retroviral infection, as outlined in PCT US97/01019, incorporated by reference.

As will be appreciated by those in the art, the type of mammalian cells used in the present invention can vary widely. Basically, any mammalian cells may be used, with mouse, rat, primate and human cells being particularly preferred, although as will be appreciated by those in the art, modifications of the system by pseudotyping allows all eukaryotic cells to be used, preferably higher eukaryotes. As is more fully described below, a screen will be set up such that the cells exhibit a selectable phenotype in the presence of a bioactive peptide. As is more fully described below, cell types implicated in a wide variety of disease conditions are particularly useful, so long as a suitable screen may be designed to allow the selection of cells that exhibit an altered phenotype as a consequence of the presence of a peptide within the cell.

Accordingly, suitable cell types include, but are not limited to, tumor cells of all types (particularly melanoma, myeloid leukemia, carcinomas of the lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell), mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoetic, neural, skin, lung, kidney, liver and myocyte stem cells (for use in screening for differentiation and de-differentiation factors), osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to, Jurkat T cells, NIH3T3 cells, CHO, Cos, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

In one embodiment, the cells may be additionally genetically engineered, that is, contain exogeneous nucleic acid other than the GPA nucleic acid.

In a preferred embodiment, the GPA proteins are expressed in bacterial systems. Bacterial expression systems are well known in the art.

A suitable bacterial promoter is any nucleic acid sequence capable of binding bacterial RNA polymerase and initiating the downstream (3') transcription of the coding sequence of the GPA protein into mRNA. A bacterial promoter has a transcription initiation region which is usually placed proximal to the 5' end of the coding sequence. This transcription initiation region typically includes an RNA polymerase binding site and a transcription initiation site. Sequences encoding metabolic pathway enzymes provide particularly useful promoter sequences. Examples include promoter sequences derived from sugar metabolizing enzymes, such as galactose, lactose and maltose, and sequences derived from biosynthetic enzymes such as tryptophan. Promoters from bacteriophage may also be used and are known in the art. In addition, synthetic promoters and hybrid promoters are also useful; for example, the tac promoter is a hybrid of the trp and lac promoter sequences. Furthermore, a bacterial promoter can include naturally occurring promoters of non-bacterial origin that have the ability to bind bacterial RNA polymerase and initiate transcription.

In addition to a functioning promoter sequence, an efficient ribosome binding site is desirable. In *E. coli*, the ribosome binding site is called the Shine-Delgarno (SD) sequence and includes an initiation codon and a sequence 3–9 nucleotides in length located 3–11 nucleotides upstream of the initiation codon.

The expression vector may also include a signal peptide sequence that provides for secretion of the GPA protein in bacteria. The signal sequence typically encodes a signal peptide comprised of hydrophobic amino acids which direct the secretion of the protein from the cell, as is well known in the art. The protein is either secreted into the growth media (gram-positive bacteria) or into the periplasmic space, located between the inner and outer membrane of the cell (gram-negative bacteria).

The bacterial expression vector may also include a selectable marker gene to allow for the selection of bacterial strains that have been transformed. Suitable selection genes include genes which render the bacteria resistant to drugs such as ampicillin, chloramphenicol, erythromycin, kanamycin, neomycin and tetracycline. Selectable markers also include biosynthetic genes, such as those in the histdine, tryptophan and leucine biosynthetic pathways.

These components are assembled into expression vectors. Expression vectors for bacteria are well known in the art, and include vectors for *Bacillus subtilis, E. coli, Streptococcus cremoris*, and *Streptococcus lividans*, among others.

The bacterial expression vectors are transformed into bacterial host cells using techniques well known in the art, such as calcium chloride treatment, electroporation, and others.

In one embodiment, GPA proteins are produced in insect cells. Expression vectors for the transformation of insect cells, and in particular, baculovirus-based expression vectors, are well known in the art.

In a preferred embodiment, GPA protein is produced in yeast cells. Yeast expression systems are well known in the art, and include expression vectors for *Saccharomyces cerevisiae, Candida albicans* and *C. maltosa, Hansenula polymorpha, Kluyveromyces fragilis* and *K. lactis, Pichia guillerimondii* and *P. pastoris, Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. Preferred promoter sequences for expression in yeast include the inducible GAL1,10 promoter, the promoters from alcohol dehydrogenase, enolase, glucokinase, glucose6-phosphate isomerase, glyceraldehyde-3-phosphate-dehydrogenase, hexokinase, phosphofructokinase, 3-phosphoglycerate mutase, pyruvate kinase, and the acid phosphatase gene. Yeast selectable markers include ADE2, HIS4, LEU2, TRP1, and ALG7, which confers resistance to tunicamycin; the neomycin phosphotransferase gene, which confers resistance to G418; and the CUP1 gene, which allows yeast to grow in the presence of copper ions.

In addition, the GPA polypeptides of the invention may be further fused to other proteins, if desired, for example to increase expression.

In one embodiment, the GPA nucleic acids, proteins and antibodies of the invention are labeled with a label other than the scaffold. By "labeled" herein is meant that a compound has at least one element, isotope or chemical compound attached to enable the detection of the compound. In general, labels fall into three classes: a) isotopic labels, which may be radioactive or heavy isotopes; b) immune labels, which may be antibodies or antigens; and c) colored or fluorescent dyes. The labels may be incorporated into the compound at any position.

Once made, the GPA proteins may be covalently modified. One type of covalent modification includes reacting targeted amino acid residues of an GPA polypeptide with an organic derivatizing agent that is capable of reacting with selected side chains or the N-or C-terminal residues of an GPA polypeptide. Derivatzation with bifunctional agents is useful, for instance, for crosslinking GPA to a water-insoluble support matrix or surface for use in the method for purifying anti-GPA antibodies or screening assays, as is more fully described below. Commonly used crosslinking agents include, e.g., 1,1-bis(diazo-acetyl)-2-phenylethane, glutaraldehyde, N-hydroxysuccinimide esters, for example, esters with 4-azidosalicylic acid, homobifunctional imidoesters, including disuccinimidyl esters such as 3,3'-dithiobis-(succinimidylpropionate), bifunctional maleimides such as bis-N-maleimido-1,8-octane and agents such as methyl-3-[(p-azidophenyl)dithio]propioimidate.

Other modifications include deamidation of glutaminyl and asparaginyl residues to the corresponding glutamyl and aspartyl residues, respectively, hydroxylabon of proline and lysine, phosphorylation of hydroxyl groups of seryl or threonyl residues, methylation of the "-amino groups of lysine, arginine, and histidine side chains [T. E. Creighton, *Proteins: Structure and Molecular Properties*, W. H. Freeman & Co., San Francisco, pp. 79–86 (1983)], acetylation of the N-terminal amine, and amidabon of any C-terminal carboxyl group.

Another type of covalent modification of the GPA polypeptide included within the scope of this invention comprises altering the native glycosylabon pattern of the polypeptide. "Altering the native glycosylation pattern" is intended for purposes herein to mean deleting one or more carbohydrate moieties found in native sequence GPA polypeptide, and/or adding one or more glycosylabon sites that are not present in the native sequence GPA polypeptide.

Addition of glycosylation sites to GPA polypeptides may be accomplished by altering the amino acid sequence thereof. The alteration may be made, for example, by the addition of, or substitution by, one or more serine or threonine residues to the native sequence GPA polypeptide (for O-linked glycosylation sites). The GPA amino acid sequence may optionally be altered through changes at the DNA level, particularly by mutating the DNA encoding the GPA polypeptide at preselected bases such that codons are generated that will translate into the desired amino acids. Another means of increasing the number of carbohydrate moieties on the GPA polypeptide is by chemical or enzymatic coupling of glycosides to the polypeptide. Such methods are described in the art, e.g., in WO 87/05330 published Sep. 11, 1987, and in Aplin and Wriston, *CRC Crit. Rev. Biochem.*, pp. 259–306 (1981).

Removal of carbohydrate moieties present on the GPA polypeptide may be accomplished chemically or enzymatically or by mutational substitution of codons encoding for amino acid residues that serve as targets for glycosylation. Chemical deglycosylation techniques are known in the art and described, for instance, by Hakimuddin, et al., *Arch. Biochem. Biophys.*, 259:52 (1987) and by Edge et al., *Anal. Biochem.*, 118:131 (1981). Enzymatic cleavage of carbohydrate moieties on polypeptides can be achieved by the use of a variety of endo-and exo-glycosidases as described by Thotakura et al., *Meth. Enzymol.*, 138:350 (1987).

Another type of covalent modification of GPA comprises linking the GPA polypeptide to one of a variety of nonproteinaceous polymers, e.g., polyethylene glycol, polypropylene glycol, or polyoxyalkylenes, in the manner set forth in U.S. Pat. Nos. 4,640,835; 4,496,689; 4,301,144; 4,670,417; 4,791,192 or 4,179,337.

GPA polypeptides of the present invention may also be modified in a way to form chimeric molecules comprising an GPA polypeptide fused to another, heterologous polypeptide or amino acid sequence. In one embodiment, such a chimeric molecule comprises a fusion of an GPA polypeptide with a tag polypeptide which provides an epitope to which an anti-tag antibody can selectively bind. The epitope tag is generally placed at the amino-or carboxyl-terminus of the GPA polypeptide. The presence of such epitope-tagged forms of an GPA polypeptide can be detected using an antibody against the tag polypeptide. Also, provision of the epitope tag enables the GPA polypeptide to be readily purified by affinity purification using an anti-tag antibody or another type of affinity matrix that binds to the epitope tag. In an alternative embodiment, the chimeric molecule may comprise a fusion of an GPA polypeptide with an immunoglobulin or a particular region of an immunoglobulin. For a bivalent form of the chimeric molecule, such a fusion could be to the Fc region of an IgG molecule.

Various tag polypeptides and their respective antibodies are well known in the art. Examples include poly-histidine (poly-his) or poly-histidine-glycine (poly-his-gly) tags; the flu HA tag polypeptide and its antibody 12CA5 [Field et al., *Mol. Cell. Biol.*, 8:2159–2165 (1988)]; the c-myc tag and the 8F9, 3C7, 6E10, G4, B7 and 9E10 antibodies thereto [Evan et al., *Molecular and Cellular Biology*, 5:3610–3616 (1985)]; and the Herpes Simplex virus glycoprotein D (gD) tag and its antibody [Paborsky et al., *Protein Engineering*, 3(6):547–553 (1990)]. Other tag polypeptides include the Flag-peptide [Hopp et al., *BioTechnoloqy*, 6:1204–1210 (1988)]; the KT3 epitope peptide [Martin et al., *Science*, 255:192–194 (1992)]; tubulin epitope peptide [Skinner et al., *J. Biol. Chem.*, 266:15163–15166 (1991)]; and the T7 gene 10 protein peptide tag [(Lutz-Freyermuth et al., *Proc. Natl. Acad. Sci. USA*, 87:6393–6397 (1990)].

In a preferred embodiment, the GPA protein is purified or isolated after expression. GPA proteins may be isolated or purified in a variety of ways known to those skilled in the art depending on what other components are present in the sample. Standard purification methods include electrophoretic, molecular, immunological and chromatographic techniques, including ion exchange, hydrophobic, affinity, and reverse-phase HPLC chromatography, and chromatofocusing. For example, the GPA protein may be purified using a standard anti-library antibody column. Ultrafiltrabon and diafiltration techniques, in conjunction with protein concentration, are also useful. For general guidance in suitable purification techniques, see Scopes, R., Protein Purification, Springer-Verlag, NY (1982). The degree of purification necessary will vary depending on the use of the GPA protein. In some instances no purification will be necessary. A preferred method for purification is outlined in the examples.

Once made, the GPA proteins and nucleic acids of the invention find use in a number of applications.

In a preferred embodiment, the GPA proteins are administered to a patent to treat a G-CSF-associated disorder.

By "G-CSF associated disorder" or "neutropenic" or "G-CSF responsive disorder" or "condition" herein is meant a disorder that can be ameliorated by the administration of a compound with a GPA protein, including, but not limited to, neutropenia associated with cancer therapies including chemotherapy and radiation therapy; radiation accidents; bone marrow transplantation; bone marrow suppression conditions, for example those associated with AIDS; myelodysplastc syndromes characterized by granulocyte functional abnormalities; severe infections; etc. In addition, treatment with the GPA proteins of the invention can be used to enhance peripheral blood progenitor cell collection.

In a preferred embodiment, a therapeutically effective dose of a GPA protein is administered to a patient. By "therapeutcally effective dose" herein is meant a dose that produces the effects for which it is administered. The exact dose will depend on the purpose of the treatment, and will be ascertainable by one skilled in the art using known techniques. In a preferred embodiment, dosages of about 5 μg/kg are used, administered either intraveneously or subcutaneously. As is known in the art, adjustments for GPA degradation, systemic versus localized delivery, and rate of new protease synthesis, as well as the age, body weight, general health, sex, diet, time of administration, drug interaction and the severity of the condition may be necessary, and will be ascertainable with routine experimentation by those skilled in the art.

A "patient" for the purposes of the present invention includes both humans and other animals, particularly mammals, and organisms. Thus the methods are applicable to both human therapy and veterinary applications. In the preferred embodiment the patient is a mammal, and in the most preferred embodiment the patient is human.

The administration of the GPA proteins of the present invention can be done in a variety of ways, including, but not limited to, orally, subcutaneously, intravenously, intranasally, transdermally, intraperitoneally, intramuscularly, intrapulmonary, vaginally, rectally, or intraocularly. In some instances, for example, in the treatment of wounds and inflammation, the GPA protein may be directly applied as a solution or spray.

The pharmaceutical compositions of the present invention comprise a GPA protein in a form suitable for administration to a patient. In the preferred embodiment, the pharmaceutical compositions are in a water soluble form, such as being present as pharmaceutically acceptable salts, which is meant to include both acid and base addition salts. "Pharmaceutcally acceptable acid addition salt" refers to those salts that retain the biological effectiveness of the free bases and that are not biologically or otherwise undesirable, formed with inorganic acids such as hydrochloric acid, hydrobromic acid, sulfuric acid, nitric acid, phosphoric acid and the like, and organic acids such as acetic acid, propionic acid, glycolic acid, pyruvic acid, oxalic acid, maleic acid, malonic acid, succinic acid, fumaric acid, tartaric acid, citric acid, benzoic acid, cinnamic acid, mandelic acid, methanesulfonic acid, ethanesulfonic acid, p-toluenesulfonic acid, salicylic acid and the like. "Pharmaceutcally acceptable base addition salts" include those derived from inorganic bases such as sodium, potassium, lithium, ammonium, calcium, magnesium, iron, zinc, copper, manganese, aluminum salts and the like.

Particularly preferred are the ammonium, potassium, sodium, calcium, and magnesium salts. Salts derived from pharmaceutically acceptable organic non-toxic bases include salts of primary, secondary, and tertiary amines, substituted amines including naturally occurring substituted amines, cyclic amines and basic ion exchange resins, such as isopropylamine, trimethylamine, diethylamine, triethylamine, tripropylamine, and ethanolamine.

The pharmaceutical compositions may also include one or more of the following: carrier proteins such as serum albumin; buffers such as NaOAc; fillers such as microcrystalline cellulose, lactose, corn and other starches; binding agents; sweeteners and other flavoring agents; coloring agents; and polyethylene glycol. Additives are well known in the art, and are used in a variety of formulations.

In a preferred embodiment, GPA proteins are administered as therapeutic agents, and can be formulated as outlined above. Similarly, GPA genes (including both the full-length sequence, partial sequences, or regulatory sequences of the GPA coding regions) can be administered in gene therapy applications, as is known in the art. These GPA genes can include antisense applications, either as gene therapy (i.e. for incorporation into the genome) or as antisense compositions, as will be appreciated by those in the art.

In a preferred embodiment, the nucleic acid encoding the GPA proteins may also be used in gene therapy. In gene therapy applications, genes are introduced into cells in order to achieve in vivo synthesis of a therapeutically effective genetic product, for example for replacement of a defective gene. "Gene therapy" includes both conventional gene therapy where a lasting effect is achieved by a single treatment, and the administration of gene therapeutic agents, which involves the one time or repeated administration of a therapeutically effective DNA or mRNA. Antisense RNAs and DNAs can be used as therapeutic agents for blocking the expression of certain genes in vivo. It has already been shown that short antisense oligonucleotides can be imported into cells where they act as inhibitors, despite their low intracellular concentrations caused by their restricted uptake by the cell membrane. (Zamecnik et al., *Proc. Natl. Acad. Sci. U.S.A.*, 83:4143–4146 [1986]). The oligonucleotides can be modified to enhance their uptake, e.g. by substituting their negatively charged phosphodiester groups by uncharged groups.

There are a variety of techniques available for introducing nucleic acids into viable cells. The techniques vary depending upon whether the nucleic acid is transferred into cultured cells in vitro, or in vivo in the cells of the intended host. Techniques suitable for the transfer of nucleic acid into mammalian cells in vitro include the use of liposomes, electroporation, microinjection, cell fusion, DEAE-dextran, the calcium phosphate precipitation method, etc. The currently preferred in vivo gene transfer techniques include transfecton with viral (typically retroviral) vectors and viral coat protein-liposome mediated transfection [Dzau et al., *Trends in Biotechnology*, 11:205–210 (1993)]. In some situations it is desirable to provide the nucleic acid source with an agent that targets the target cells, such as an antibody specific for a cell surface membrane protein or the target cell, a ligand for a receptor on the target cell, etc. Where liposomes are employed, proteins which bind to a cell surface membrane protein associated with endocytosis may be used for targeting and/or to facilitate uptake, e.g. capsid proteins or fragments thereof tropic for a particular cell type, antibodies for proteins which undergo internalization in cycling, proteins that target intracellular localization and enhance intracellular half-life. The technique of receptor-mediated endocytosis is described, for example, by Wu et al., *J. Biol. Chem.*, 262:4429–4432 (1987); and Wagner et al., *Proc. Natl. Acad. Sci. U.S.A.*, 87:3410–3414 (1990). For review of gene marking and gene therapy protocols see Anderson et al., *Science*, 256:808–813 (1992).

In a preferred embodiment, GPA genes are administered as DNA vaccines, either single genes or combinations of GPA genes. Naked DNA vaccines are generally known in the art. Brower, Nature Biotechnology, 16:1304–1305 (1998). Methods for the use of genes as DNA vaccines are well known to one of ordinary skill in the art, and include placing a GPA gene or portion of a GPA gene under the control of a promoter for expression in a GPA patient. The GPA gene used for DNA vaccines can encode full-length GPA proteins, but more preferably encodes portions of the GPA proteins including peptides derived from the GPA protein. In a preferred embodiment a patient is immunized with a DNA vaccine comprising a plurality of nucleotide sequences derived from a GPA gene. Similarly, it is possible to immunize a patient with a plurality of GPA genes or portions thereof as defined herein. Without being bound by theory, expression of the polypeptide encoded by the DNA vaccine, cytotoxic T-cells, helper T-cells and antibodies are induced which recognize and destroy or eliminate cells expressing GPA proteins.

In a preferred embodiment, the DNA vaccines include a gene encoding an adjuvant molecule with the DNA vaccine. Such adjuvant molecules include cytokines that increase the immunogenic response to the GPA polypeptide encoded by the DNA vaccine. Additional or alternative adjuvants are known to those of ordinary skill in the art and find use in the invention.

The following examples serve to more fully describe the manner of using the above-described invention, as well as to set forth the best modes contemplated for carrying out various aspects of the invention. It is understood that these examples in no way serve to limit the true scope of this invention, but rather are presented for illustrative purposes. All references cited herein are incorporated by reference in their entirety.

EXAMPLES

Design and Characterization of Novel GPA Proteins

Protein Design

Summary: Sequences for novel granulopoietic proteins (GPA proteins) were designed by simultaneously optimizing residues in the buried core of the protein using Protein Design Automation (PDA) as described in WO98/47089 and U.S. Ser. No. 09/127,926, both of which are expressly incorporated by reference in their entirety. Several core designs were completed, with 25–34 residues considered, corresponding to 10^{27}–10^{28} sequence possibilities. Residues unexposed to solvent were designed in order to minimize changes to the molecular surface and to limit the potential for antigenicity of designed novel protein analogues. Calculations required from 12–24 hours on 16 Silicon Graphics R10000 CPU's. The global optimum sequence from each design was selected for characterization. From 10–14 residues were changed from hG-CSF in the designed proteins, out of 174 residues total. Additional designs were done where 14–24 boundary positions were optimized resulting in 12–20 mutated residues. These designs were repeated using the optimal sequence obtained from one of the core designs as the template structure, again producing optimal sequences with from 12–20 mutations. Only the global optimum sequences were selected for experimental study because of the high stringency of PDA and the very low false positive rate.

Computational Protocols

Template structure preparation: The template structure was produced using homology modeling. The crystal structure of bovine G-CSF (PDB record 1 bgc) was used as the starting point for modeling since the crystal structure of human G-CSF is at lower resolution and is missing key fragments including a restraining disulfide bond between positions 64 and 74. Bovine G-CSF also serves as a good model for human G-CSF since the sequences are the same length and 142 out of 174 amino acids are identical (81%). The 32 residues that differ in the bovine sequence were replaced with the human residues for those positions and the conformations of the replaced side chains were optimized using PDA. The optimization was initially done on all the replaced residues except position 167; typical PDA parameters were used (the van der Waals scale factor was set to 0.9, the H-bond potential well-depth was set to 8.0 kcal/mol, and the solvation potential was calculated using type 2 solvation with a nonpolar burial energy of 0.048 kcal/mol and a nonpolar exposure multiplication factor of 1.6). For position 167, the Gly in bovine G-CSF was replaced with the human residue for this position (Val).

However, due to steric constraints between position 167 and the disulfide bond between positions 64 and 74, the Val at this position was optimized using less restrictive steric constraints (PDA was run using a van der Waals scale factor of 0.7 instead of the typical value of 0.9). The entire structure was then minimized for 50 steps using conjugate gradient minimization and the Dreiding II force field. This minimized structure was used as the template for all the designs.

Design strategies: Core residues were selected for design since optimization of these positions can improve stability, although stabilization has been obtained from modifications at other sites as well. Core designs also minimize changes to the molecular surface and thus limit the designed protein's potential for antigenicity. PDA calculations were run on three core designs; core3 had 34 core positions that were allowed to vary, core4 had 26, and core4v had 25 (see FIG. 2). The core3 variable positions were selected from the entire length of the helices, while core4 and core4v's variable positions were selected from the interior (not at the ends) of the helices. Only hydrophobic amino acids were allowed at the variable core positions. These included Ala, Val, Phe, Ile, Leu, Tyr and Trp. Gly was also allowed for the variable positions that had Gly in the bovine wild type structure (positions 28, 149, 150, and 167). Met and Pro were not allowed.

Two boundary designs were also done; bndry4_2 had 24 variable boundary residues, and bndry4_AD had 14 (see FIG. 2). The bndry4_AD design was restricted to boundary residues on the outer two helices (A and D) since initial calculations suggested that the most pronounced changes in helical propensity result from modifications at these locations, and we anticipated that improvements in helical propensity might lead to improved stability. Two additional boundary designs were done (bndry4_2_core4 and bndry4_AD_core4) which allowed the same boundary positions to vary but used the optimal sequence from the core4 design as the template. That is, these designs were required to keep the 10 core mutations (amino acid and conformation) that resulted from the core4 PDA calculations (see FIG. 3). The boundary designs allowed the following amino acids at the variable positions: Ala, Val, Leu, Ile, Asp, Asn, Glu, Gln, Lys, Ser, Thr, and Hsp (a protonated His). Met, Pro, Cys, Gly, Arg, and the aromatics Trp, Tyr, and Phe were not allowed.

PDA Calculations

The PDA calculations for all the designs were run using the a2h1p0 rotamer library. This library is based on the backbone-dependent rotamer library of Dunbrack and Karplus (Dunbrack and Karplus, 1993) but includes more rotamers for the aromatic and hydrophobic amino acids; X_1 and X2 angle values of rotamers for all the aromatic amino acids and X_1 angle values for all the other hydrophobic amino acids were expanded ±1 standard deviation about the mean value reported in the Dunbrack and Karplus library. Typical PDA parameters were used: the van der Waals scale factor was set to 0.9, the H-bond potential well-depth was set to 8.0 kcal/mol, the solvation potential was calculated using type 2 solvation with a nonpolar burial energy of 0.048 kcal/mol and a nonpolar exposure multiplication factor of 1.6, and the secondary structure scale factor was set to 0.0 (secondary structure propensities were not considered). Calculations required from 12–24 hours on 16 Silicon Graphics R10000 CPU's.

Optimal Sequences

The optimal sequence selected by PDA for each of the designs is shown in FIG. 3. In the core designs, from 10 to 14 residues were changed compared to wild type, while the boundary designs produced 20 mutations for bndry4_2 (all four helices designed) and 12 mutations for bndry4_AD (only A and D helices designed). Including the core4 mutants in the template resulted in the same number of boundary mutations (20 for bndry4_core4; 12 for bndry4_AD_core4), but different amino acids were selected at some of the mutated positions.

Monte Carlo Analysis

Monte Carlo analysis of the sequences produced by PDA shows the ground state (optimal) amino acid and amino acids allowed for each variable position and their frequencies of occurrence (see FIGS. 4 through 10).

Cloning and Expression

Summary: A gene for met hG-CSF was synthesized from partially overlapping oligonucleotides (approximately 100 bases) that were extended and PCR amplified; see FIG. 1B. Codon usage was optimized for *E. coli* and several restriction sites were incorporated to ease future cloning. These partial genes were cloned into a vector and transformed into *E. coli* for sequencing. Several of these gene fragments were then cloned into adjacent positions in an expression vector (pET17 or pET21) to form the full length gene for met hG-CSF (528 bases) and transformed into *E coli* for expression. Protein was expressed in *E. coli* in insoluble inclusion bodies (data not shown) and its identity was confirmed by immunoblot of SDS-PAGE using a commercial Mab against hG-CSF. A similar strategy was followed for all of the novel GPA proteins and all were expressed (data not shown).

Cloning

To clone the gene, pairs of partially complementary oligonucleotides were synthesized and annealed by heating to 70° C. for 10 min and cooling to room temperature. The overlapping oligonucleotides (100 mers) were extended using Klenow fragment for 1 hour at 37° C. These extended oligonucleotides were then used as templates for PCR with primers complementary to the terminal 20 nucleotides of each end. PCR products were cloned into the vector pCR-Blunt (Invitrogen) according to the manufacturer's recommendations, and transformed into Gibco-BRL Subcloning Efficiency *E. coli* DH5α cells. The DNAs from several colonies were isolated using a Qiagen Miniprep Spin Kit, and sequenced by an Applied Biosystems 377XL automated flourescent DNA sequencer.

Expression

To express the protein, sequenced genes were subcloned between the NdeI and XhoI sites of Novagen's pET21a (+) vector and transformed into *E. coli* BL21 (DE3) cells. Protein expression was induced by growing the *E. coli* cells in Circlegrow media (Bio 101) with shaking at 37° C. to a density of 0.5 OD_{550}. IPTG was then added to a final concentration of 1 mM, and growth was allowed to continue for a further 3 hours. The expressed protein incorporated a Met at the N-terminus; our numbering begins with the next residue, a Thr.

To confirm expression of the protein, 10 μl samples were removed prior to addition of IPTG and at the end of the three hour incubation. These samples were electrophoresed through a 15% SDS-polyacrylamide gel and stained with Coomassie blue R-250. Expression of protein with the expected molecular weight could readily be observed. Confirmation that the protein was GCSF was obtained by immunoblot analysis using monoclonal antibodies directed to either the N-terminal 20 amino acids or the C-terminal 18 amino acids (Santa Cruz Biotechnology).

Isolation and Purification

Summary: Protein was isolated by solubilizing the inclusion bodies in detergent and refolding the protein in the presence of $CuSO_4$ to promote formation of native disulfide bonds. The solubilized protein mixture was loaded onto a size exclusion column to separate monomeric protein from aggregates and contaminants from the preparation. Fractions containing monomeric met hG-CSF were collected and assessed for purity by reversed phase HPLC. Greater than 95% purity was confirmed. The designed GPA proteins eluted slightly later than wildtype met hG-CSF.

HPLC purification: The mixture was directly loaded onto the size exclusion column (10 mm×300 mm loaded with superdex prep 75 resin purchased from Pharmacia) and eluted at a flow rate of 0.8 ml/min using the column buffer (100 mM Na_2SO_4, 50 mM Tris, pH 7.5). The peaks are monitered by UV detector at dual wavelengths of 214 and 280 nm. Albumin, carbonic anhydrate, cytochrome C and aprotinin were used to calibrate the molecular size of proteins versus elution time. The monomeric peak that elutes around the expected elution time for each protein was collected and the buffer was exchanged into 10mM NaOAc at pH 4 for biophysical characterization. For long term storage, a buffer of 5% sorbitol, 0.004% Tween 80, and 10 mM NaOAc at pH 4 was used. The proteins were >98% pure as judged by reversed phase HPLC on a C4 column (3.9 mm×150 mm) with linear acetonitrile-water gradient containing 0.1% TFE.

Isolation and refolding from inclusion bodies: To isolate inclusion bodies, the *E. coli* cells were pelleted by centrifugation at 8000 rpm in a Beckman J2-17 rotor. The cells were re-suspended in 50 mM Tris.HCl pH8.0, 10 mM $MgCl_2$ at 5 mls per gram of pelleted cells. Lysozyme was added to a final concentration of 0.1 mg/ml, and the cells were incubated at 30° C. for 30 min. The cells were then rapidly frozen and thawed, and DNase 1 was added to a concentration of 10 μg/ml. After incubation at 37° C. for 30 minutes, the inclusion bodies were isolated by centrifugation at 12 000 rpm for 30 min and washed twice with 50 mM Tris.HCl pH8.0, 10 mM $MgCl_2$.

The protein precipitate was washed and fully solubilized in 2% sarkosyl, 50 mM Tris, pH 8.0. $CuSO_4$ was then added into the mixture to reach a concentration of 20 uM. The mixture was stirred for 8–10 hours to refold the proteins by forming disulfide bonds under air oxidation.

Spectroscopic Characterization

Summary: Protein structure was assessed by circular dichroism (CD). The CD spectra for met hG-CSF and the GPA proteins tested were nearly identical to each other and to published spectra of met hG-CSF. These spectra indicate highly similar secondary structure and tertiary folds for the GPA proteins and met hG-CSF. Thermal stability was assessed by monitoring the temperature dependence of the CD signal at 222 nm, a wavelength diagnostic for helical protein structure. The thermal stabilities of the proteins are shown in FIG. **13**, with core4 approximately 10° C. more stable than met hG-CSF and core3 and core4v having very similar thermal stabilities to met hG-CSF. As in previously published PDA designed proteins, the origin of the increased stability likely results from an improved balance between packing interactions and hydrophobic burial of side chains. The thermal stabilities of three additional GPA proteins (sm0, fm3 and fm4) derived by reverting some of the core mutant positions to wild type are shown in FIG. **16**.

Spectroscopic Characterization

The concentrations of the proteins were determined by UV spectroscopy at 280 nm using the extinction coefficients shown in FIG. **16**. CD spectra were measured on an Aviv 202DS spectrometer equipped with a Peltier temperature control unit. The ellipticity was calibrated with (+)-10-camphorsulfonic acid. The thermal transition curves were recorded at 222 nm in a buffer of 10 mM NaOAc at pH 4.0 every 2.5° C. with an averaging time of 5 s and an equilibration time of 3 min. The melting temperature (T_m) value of each protein was derived from the derivative curve of the ellipticity at 222 nm vs. temperature. The T_m values were reproducible to within 1° for the same protein at the concentrations used (~0.1 mg/ml). Thermal denaturation curves are shown in FIG. **13**. The T_m's for core4, core4v and core3 and three proteins derived from them (sm0, fm4 and fm7) are shown in FIG. **16**.

In Vitro Biological Activity

Summary: FIG. **14** shows the dose response curves for met hG-CSF and three GPA proteins. Mouse leukocytes were transfected with human G-CSF receptor, making leukocyte proliferation dependent on G-CSF signaling activity via the G-CSF receptor. Leukocyte proliferation is measured by incorporation of brominated uracil (BrdU) measured by ELISA. GPA protein granulopoietic activity is measured by quantifying cell proliferation as a function of protein concentration. Two hG-CSF samples were also tested, one produced as described herein and a commercially available hG-CSF from R&D Systems. Dose response curves were very similar for all of the proteins tested, except for core4, which showed approximately two times the potency of met hG-CSF. FIG. **15** shows the appearance of a typical 96 well plate ELISA of control samples with met hG-CSF. The statistical analysis of the dose response assay (8 replicates) shows that core4 was highly significantly more potent than the other GPA proteins and met hG-CSF. The origin of this effect is unclear, and could be from increased affinity for the receptor, increased stability of core4 under cell culture assay conditions, or a combination.

Cell culture: The cells used in the proliferation assay were Ba/F3 (murine lymphoid) cells stably transfected with the gene encoding the human Class 1 GCSF receptor (a kind gift from Dr. Belinda Avalos, Ohio State University). These cells were maintained in RPM1 medium 1640 (Gibco-BRL) at 5% CO_2, 37° C. in high humidity. They were passaged every 2–3 days by a 1 in 10 dilution into fresh media.

Cell proliferation assay: Cell proliferation in response to GCSF was detected by 5-bromo-2'-deoxyuridine (BrdU) incorporation quantified by a BrdU-specific ELISA kit as described by the manufacturer (Boehringer Mannheim). Briefly, 1×10^5 to 1×10^6 Ba/F3 cells/ml are incubated with varying amounts of GCSF (1×10^2 pg/ml to 1×10^5 pg/ml) for 42 hrs before the addition of 10 μM BrdU. After further incubation of 22 hrs, the cells are lysed and the DNA denatured using FixDenat (Boehringer Mannheim). Incorporation of BrdU into DNA was then quantified with an ELISA that utilizes a peroxidase-conjugated monoclonal antibody against BrdU. Peroxidase activity was measured at 450 nm by a BioRad Model 550 microtitre plate reader. Typically, each experiment contained 8 replicates spread over 4 plates. Data was analyzed by Kaleidagraph (Synergy Software) and Statistica (Statsoft).

Storage Stability

The storage stability of core4 was assessed by incubation at both 37 and 50° C. under solution conditions identical in composition to that used in the commercial formulation of Neupogen. Accelerated degradation was followed by observing the disappearance of monomeric protein with size exclusion chromatography, since aggregation is the predominant mechanism of inactivation of G-CSF. Even under optimized formulation conditions, core4 is significantly more stable than met hG-CSF (FIG. **15**).

SEQUENCE LISTING

```
<160> NUMBER OF SEQ ID NOS: 18

<210> SEQ ID NO 1
<211> LENGTH: 526
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 1

atgactccat taggtccagc ttcctctctg ccgcaaagct tcctgctgaa atgcctggaa      60

caggttcgta aaatccaggg tgatggtgct gctctgcagg aaaaactgtg cgctacctac     120

aaactgtgcc atccggaaga actggttctg ctgggtcact ccctgggtat cccgtgggcg     180

ccgctgagct cctgcccgag ccaggctctg cagctggctg gttgcctgtc ccaattgcac     240

agcggccttt tcctgtacca gggtctgctg caagctctgg aaggtactcc ccggaactgg     300

gtccgaccct ggacactctg cagctggacg tcgctgactt cgctaccacc atctggcagc     360

agatggaaga actgggtatg gctccggctc tgcagccgac ccagggtgct atgccggctt     420

tcgttccgct ttccagcgtc gcgcaggtgg cgttctggtt gctagccacc tgcagagctt     480

cctggaagtt tcctaccgtg ttctgcgtca cctggctcag ccgtga                    526


<210> SEQ ID NO 2
<211> LENGTH: 174
<212> TYPE: PRT
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 2

Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu Lys
1               5                   10                  15

Cys Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Gly Ala Ala Leu Gln
            20                  25                  30

Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu Val
        35                  40                  45
```

-continued

```
Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser Cys
    50                  55                  60

Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu His Ser
65                  70                  75                  80

Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile Ser
                85                  90                  95

Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala Asp
            100                 105                 110

Phe Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala Pro
        115                 120                 125

Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala Phe
    130                 135                 140

Gln Arg Arg Ala Gly Gly Val Leu Val Ala Ser His Leu Gln Ser Phe
145                 150                 155                 160

Leu Glu Val Ser Tyr Arg Val Leu Arg His Leu Ala Gln Pro
                165                 170


<210> SEQ ID NO 3
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 3

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Ile Leu
-1  1               5                   10                  15

Lys Cys Leu Glu Leu Val Arg Lys Ile Gln Gly Glu Gly Ala Ala Leu
                20                  25                  30

Ile Glu Ile Leu Cys Ala Lys Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu Leu
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Lys Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Val Gly Pro Ile Leu Asp Thr Leu Ile Leu Glu Val Ala
                100                 105                 110

Asp Phe Ala Thr Ile Ile Trp Gln Leu Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Lys Glu Asp Gly Gly Val Leu Val Ala Ile Leu Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ala Tyr Arg Val Leu Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 4
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
```

-continued

```
<222> LOCATION: (2)..()

<400> SEQUENCE: 4

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Ile Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Leu Val Arg Lys Ile Gln Gly Glu Ala Ala Ala Leu
                20                  25                  30

Leu Glu Glu Leu Cys Ala His Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe Leu
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Leu Gln Lys Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Lys Val Asp Thr Leu Ile Leu Glu Ile Ala
                100                 105                 110

Asp Leu Ala Thr Ile Ile Trp Gln Leu Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Lys Glu Asp Gly Gly Ile Leu Ile Ala Ile Leu Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ala Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 5
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 5

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Ile Leu
-1  1               5                   10                  15

Lys Cys Leu Glu Leu Val Arg Lys Ile Gln Gly Glu Gly Ala Ala Leu
                20                  25                  30

Ile Glu Glu Leu Cys Ala His Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Phe Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Lys Glu Thr Gly Gly Val Leu Val Ala Ile Leu Leu Gln Ser
    145                 150                 155
```

-continued

```
Phe Leu Glu Val Ala Tyr Arg Val Leu Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 6
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 6

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Ile Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Leu Val Arg Lys Ile Gln Gly Glu Ala Ala Ala Leu
                20                  25                  30

Leu Glu Glu Leu Cys Ala His Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Val Asp Thr Leu Gln Leu Asp Ile Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Lys Glu Asp Gly Gly Ile Leu Ile Ala Ile Leu Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ala Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 7
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 7

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
```

-continued

```
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Val Asp Thr Leu Gln Leu Asp Ile Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 8
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 8

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Gln Ile Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Ile Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Ile Gln Ser
    145                 150                 155

Trp Phe Glu Val Ser Tyr Arg Ala Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 9
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 9

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15
```

-continued

```
Lys Leu Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Ile Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Phe Gln Ala Phe Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 10
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 10

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Ala Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Phe Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Val Leu Val Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Leu Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 11
<211> LENGTH: 175
```

-continued

```
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 11

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Ala Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Phe Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Leu Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 12
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 12

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125
```

```
Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Val Leu Val Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 13
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 13

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 14
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 14

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Leu Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Ala Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
```

-continued

```
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Phe His
    65                  70                  75

Ser Gly Leu Phe Leu Phe Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Leu Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Ile Leu Ile Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Phe Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 15
<211> LENGTH: 175
<212> TYPE: PRT
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<221> NAME/KEY: mat_peptide
<222> LOCATION: (2)..()

<400> SEQUENCE: 15

Met Thr Pro Leu Gly Pro Ala Ser Ser Leu Pro Gln Ser Phe Leu Leu
-1  1               5                   10                  15

Lys Cys Leu Glu Gln Val Arg Lys Ile Gln Gly Asp Gly Ala Ala Leu
                20                  25                  30

Gln Glu Lys Leu Cys Ala Thr Tyr Lys Leu Cys His Pro Glu Glu Leu
            35                  40                  45

Val Leu Leu Gly His Ser Leu Gly Ile Pro Trp Ala Pro Leu Ser Ser
        50                  55                  60

Cys Pro Ser Gln Ala Leu Gln Leu Ala Gly Cys Leu Ser Gln Leu His
    65                  70                  75

Ser Gly Leu Phe Leu Tyr Gln Gly Leu Leu Gln Ala Leu Glu Gly Ile
80                  85                  90                  95

Ser Pro Glu Leu Gly Pro Thr Leu Asp Thr Leu Gln Leu Asp Val Ala
                100                 105                 110

Asp Phe Ala Thr Thr Ile Trp Gln Gln Met Glu Glu Leu Gly Met Ala
            115                 120                 125

Pro Ala Leu Gln Pro Thr Gln Gly Ala Met Pro Ala Phe Ala Ser Ala
        130                 135                 140

Phe Gln Arg Arg Ala Gly Gly Val Leu Val Ala Ser His Leu Gln Ser
    145                 150                 155

Phe Leu Glu Val Ser Tyr Arg Val Leu Arg His Leu Ala Gln Pro
160                 165                 170


<210> SEQ ID NO 16
<211> LENGTH: 528
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 16

atgactccat taggtccagc ttcctctctg ccgcaaagct tcctgctgaa actgctggaa      60
```

-continued

```
caggttcgta aaatccaggg tgatgcagct gctctgcagg aaaaaatctg cgctacctac    120

aaactgtgcc atccggaaga actggttctg ctgggtcact ccctgggtat cccgtgggcg    180

ccgctgagct cctgcccgag ccaggctctg cagctggctg gttgcctgtc ccaattccac    240

agcggccttt tcctgttcca gggtctgttc caggctttcg aaggtatctc cccggaactg    300

ggtccgaccc tggacactct gcagctggac gtcgctgacc tggctaccac catctggcag    360

cagatggaag aactgggtat ggctccggct ctgcagccga cccagggtgc tatgccggct    420

ttcgcttccc ctttccagcg tcgcgcaggt ggcatcctga tcgctagcca cctgcagagc    480

ttcctggaag tttcctaccg tgttttccgt cacctggctc agccgtga                 528


<210> SEQ ID NO 17
<211> LENGTH: 528
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 17

atgactccat taggtccagc ttcctctctg ccgcaaagct tcctgctgaa actgctggaa     60

caggttcgta aaatccaggg tgatgcagct gctctgcagg aaaaactgtg cgctacctac    120

aaactgtgcc atccggaaga actggttctg ctgggtcact ccctgggtat cccgtgggcg    180

ccgctgagct cctgcccgag ccaggctctg cagctggctg gttgcctgtc ccaattccac    240

agcggccttt tcctgttcca gggtctgctg caagctctgg aaggtatctc cccggaactg    300

ggtccgaccg ttgacactct gcagctggac atcgctgacc tggctaccac catctggcag    360

cagatggaag aactgggtat ggctccggct ctgcagccga cccagggtgc tatgccggct    420

ttcgcttccg ctttccagcg tcgcgcaggt ggcatcctga tcgctagcca cctgcagagc    480

ttcctggaag tttcctaccg tgttttccgt cacctggctc agccgtga                 528


<210> SEQ ID NO 18
<211> LENGTH: 528
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 18

atgactccat taggtccagc ttcctctctg ccgcaaagct tcctgctgaa actgctggaa     60

cagatccgta aaatccaggg tgatgcagct gctctgcagg aaaaactgtg cgctacctac    120

aaactgtgcc atccggaaga actggttctg ctgggtcact ccctgggtat cccgtgggcg    180

ccgctgagct cctgcccgag ccaggctctg cagctggctg gttgcctgtc ccaattccac    240

agcggccttt tcctgttcca gggtctgctg caagctctgg aaggtatctc cccggaactg    300

ggtccgaccc tggacactct gcagctggac atcgctgacc tggctaccac catctggcag    360

cagatggaag aactgggtat ggctccggct ctgcagccga cccagggtgc tatgccggct    420

ttcgcttccg ctttccagcg tcgcgcaggt ggcatcctga tcgctagcca catccagagc    480

tggttcgaag tttcctaccg tgctttccgt cacctggctc agccgtga                 528
```

We claim:

1. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the amino acid residues at positions selected from 14, 17, 20, 21, 24, 27, 28, 31, 32, 34, 35, 38, 78, 79, 85, 89, 91, 92, 99, 102, 103, 107, 109, 110, 113, 116, 120, 145, 146, 147, 148, 151, 153, 155, 156, 157, 160, 161, 163, 164, 167, 168 and 170.

2. A non-naturally occurring GPA protein according to claim 1 wherein said GPA protein has at least 10 amino acid substitutions.

3. A non-naturally occurring GPA protein according to claim 2 wherein 10 of said substitutions are at positions 17, 28, 78, 85, 103, 110, 113, 151, 153 and 168.

4. A non-naturally occurring GPA protein according to claim 3 wherein said substitutions are 17L, 28A, 78F, 85F, 103V, 110I, 113L, 151I, 153I and 168F (SEQ ID NO: 7).

5. A non-naturally occurring GPA protein according to claim 1, wherein at least five of said substitutions are selected from the amino acid residues at positions selected from 14, 20, 27, 32, 34, 38, 79, 91, 102, 107, 109, 116, 120, 146, 147, 148, 155, 156 and 163.

6. A GPA protein according to claim 5 wherein said substitutions are 14I, 20L, 27E, 32L, 34E, 38H, 79L, 91K, 102K, 107I, 109E, 116I, 120L, 146K, 147E, 148D, 155I, 156L and 163A, (SEQ ID NO: 18).

7. A recombinant nucleic acid encoding the non-naturally occurring GPA protein of claim 1.

8. An expression vector comprising the recombinant nucleic acid of claim 7.

9. A host cell comprising the expression vector of claim 8.

10. A host cell comprising the recombinant nucleic acid of claim 1.

11. A method of producing a non-naturally occurring GPA protein comprising culturing the host cell of claim 10 under conditions suitable for expression of said nucleic acid.

12. The method according to claim 11 further comprising recovering said GPA protein.

13. A pharmaceutical composition comprising a GPA protein according to claim 1 and a pharmaceutical carrier.

14. A non-naturally occurring GPA protein according to claim 1, wherein at least five of said substitutions comprises substitutions at positions selected from 14, 17, 20, 27, 28, 32, 34, 35, 38, 78, 79, 85, 89, 9, 92, 102, 103, 107, 109, 110, 113, 116, 120, 146, 147, 148, 151, 153, 155, 156, 164, 167, and 168.

15. A non-naturally occurring GPA protein according to claim 14 wherein said substitutions are 14I, 17L, 20L, 27E, 28A, 32L, 34E, 38H, 78F, 79L, 85F, 91K, 102K, 103V, 107I, 109E, 110I, 113L, 116I, 120L, 146K, 147E, 148D, 151I, 153I, 155I, 156L, 164A and 168F (SEQ ID NO: 4).

16. A non-naturally occurring GPA protein according to claim 14 wherein said substitutions are 17L, 28A, 35I, 78F, 85F, 89F, 92F, 113L, 151I, 153I, and 168F (SEQ ID NO: 9).

17. A non-naturally occurring GPA protein according to claim 14 wherein said substitutions are 17L, 21I, 28A, 78F, 85F, 110I, 113L, 151I, 153I, 157I, 160W, 161F, 167A, and 168F (SEQ ID NO: 8).

18. A non-naturally occurring GPA protein according to claim 14 wherein said substitutions are 17L, 28A, 78F, 85F, 113L, 151I, 153I, and 168F (SEQ ID NO. 14).

19. A non-naturally occurring GPA protein comprising the amino acid sequence of SEQ ID NO: 10.

20. A non-naturally occurring GPA polypeptide comprising the amino acid sequence of SEQ ID NO: 11.

21. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the amino acid residues at positions selected from 17L, 17I; 21V, 21I: 24V, 24I; 28A, 28L; 31V, 31L; 78F; 85F, 85Y; 89L, 89F; 103V, 103L, 103I; 110V, 110L, 110I; 113L; 151I; 153V, 153I; 157L, 157I; 160F, 160W; 161L, 161F; and 168F.

22. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the group consisting of 17L, 17V, 17I, 21V, 21I, 21F; 24I, 24V; 28A, 28L; 31L, 31A, 31V, 31I; 78F, 78V; 82L, 82F; 85F, 85V, 85I, 85Y; 89L, 89F, 89W; 103V, 103A, 103L, 103I; 106L, 106V; 110I, 110V, 110L; 113L; 151I, 153I, 153V; 157L, 157V, 157I; 160F, 160W; 161L, 161F; 167A; and 168F.

23. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the group consisting of 17L, 17V, 17I; 21V, 21A, 21I, 21F, 21Y; 24I, 24A, 24V, 24L; 28A, 28L; 31L, 31V; 35I, 35V; 78F, 78A, 78V, 78L, 78I, 78Y; 82L, 82A, 82F; 85F, 85W; 89F, 89L, 89W; 92F; 103L; 103L, 103I; 106L, 106V; 110V, 110A, 110L, 110I; 113L, 113A, 113F; 117I, 117A, 117V, 117L, 117F, 117W; 151I, 153I; 157L, 157A, 157V, 157I; 160F, 160W; 161L, 161A, 161V, 161F; and 168F.

24. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the group consisting of 14I; 20L; 27E; 32I; 34K, 34I, 34F; 38V, 38I, 38E, 38K; 79L; 91K; 99V, 99L; 102L, 102I; 107I; 109E, 109V; 116I, 116L, 116K; 120L; 145Q, 145E; 146K, 146Q; 147E; 148T, 148A, 148D; 155I; 156L; 164A; 170H, 170L, 170E, and 170Q.

25. A non-naturally occurring OPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions are selected from the group consisting of 14I, 14L; 20L; 27E, 27S; 32L, 32V, 32I; 34E, 34Q, 34K; 38H, 38V, 38I, 38E, 38K; 79L; 91K; 99L, 99E; 102K, 102T, 102V, 102L, 102I, 102E, 102Q; 107I, 107V, 107L; 109E, 109V, 109D, 109Q; 116I, 116V, 116L, 116E, 116K; 120L; 145Q, 145E; 146L, 146Q; 147E, 147K; 148D, 148A, 148T; 155I; 156L; 164A; 170?, 170D, 170L, 170E, 170Q, and 170K.

26. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions arc selected from the group consisting of 14I, 14L; 20L; 27E; 32T; 34E, 34I, 34Q, 34K; 38V, 38I, 38H, 38E, 38K; 145Q, 145E; 146L, 146Q; 147E; 148T, 148A, 148D; 155I; 156L; 164A; 170H, 170D, 170L, 170E, 170Q, and 170K.

27. A non-naturally occurring GPA protein comprising at least five amino acid substitutions as compared to hG-CSF protein, wherein at least five of said substitutions arc selected from the group consisting of 14I, 14L; 20L; 27E; 32L, 32V, 32I; 34E, 34Q, 34K; 38V, 38I, 38H, 38E, 38K; 145Q, 145E; 146L, 146Q; 147E; 148D, 148A, 148T; 155I; 156L; 164S; 170H, 170L, 170E, and 170Q.

* * * * *

US006514729B1

(12) **United States Patent**
Bentzien

(10) **Patent No.: US 6,514,729 B1**
(45) **Date of Patent: Feb. 4, 2003**

(54) **RECOMBINANT INTERFERON-BETA MUTEINS**

(75) Inventor: **Jörg Bentzien**, Pasadena, CA (US)

(73) Assignee: **Xencor, Inc.**, Monrovia, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/569,722**

(22) Filed: **May 11, 2000**

Related U.S. Application Data

(60) Provisional application No. 60/133,785, filed on May 12, 1999.

(51) **Int. Cl.**^{7} **C12N 15/00**

(52) **U.S. Cl.** **435/69.51**; 435/442; 435/7.1; 530/351; 424/85.6

(58) **Field of Search** 435/69.51, 7.1, 435/442; 530/351; 424/85.6

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,518,584 A		5/1985	Mark et al.	424/85
4,588,585 A		5/1986	Mark et al.	424/85
4,737,462 A		4/1988	Mark et al.	435/253
4,738,844 A		4/1988	Bell et al.	424/85
4,738,845 A		4/1988	Bell et al.	424/85
4,753,795 A		6/1988	Bell et al.	424/85
4,769,233 A		9/1988	Bell et al.	424/85
4,793,995 A		12/1988	Bell et al.	424/85.6
4,885,166 A	*	12/1989	Meyer et al.	424/85.7
4,914,033 A		4/1990	Bell et al.	435/252.3
4,959,314 A		9/1990	Mark et al.	435/69.1
5,183,746 A		2/1993	Shaked et al.	435/69.51
5,376,567 A		12/1994	McCormick et al.	435/320.1
5,545,723 A		8/1996	Goelz et al.	424/85.6
5,730,969 A		3/1998	Hora et al.	424/85.1
5,814,485 A		9/1998	Dorin et al.	435/69.51
5,869,603 A		2/1999	Hoeprich, Jr.	530/328

FOREIGN PATENT DOCUMENTS

WO	98/47089	10/1998
WO	98/48018	10/1998

OTHER PUBLICATIONS

Shepard et al. A single amino acid change in IFN-b1 abolishes its antiviral activity. 1981. Nature, 294:563–565.*

Runkel et al. Systematic mutational mapping of sites on human IFN-b-1a that are important for receptor binding and functional activity. 2000. Biochemistry, 39:2538–51.*

Lengyel, "Biochemistry of Interferons and Their Actions", *Annu. Rev. Biochem.* 51:251–82 (1982).

Gresser and Tovey, "Antitumor Effects of Interferon", *Biochim. Biophys. Acta* 516(2):231–47 (1978).

Gresser et al., "Effect of Interferon Treatment of L1210 Cells in vitro on Tumour and Colony Formation", *Nature New Biol.* 231(18):20–1 (1971).

Dolei et al., "Interferon Effects on Friend Leukaemia Cells. I. Expression of Virus and Erythroid markers in Untreated and Dimethyl Supphoxide-treated Cells", *J. Gen. Virol.* 46(1):227–36 (1980).

Gresser, "On the Varied Biologic Effects of Interferon", *Cell Immunol* 34(2):406–15 (1977).

Stewart, "Interferon Nomenclature Recommendations", *J. Infect. Dis.* 142(4):643 (1980).

Knight, "Interferon: Purification and initial characterization from human diploid cells", *Proc. Natl. Acad. Sci. U.S.A.* 73(2):520–523 (1976).

Li et al., "Cooperative Binding of Stat1-2 heterodimers and ISGF3 to tandem DNA elements," *Biochemie* 80(8–9):703–10 (1998).

Nadeau et al., "The Proximal Tyrosines of the Cytoplasmic Domain of the β Chain of the Type I Interferon Receptor Are Essential for Signal Transducer and Activator of Transcription (Stat) 2 Activation", *J. Biol. Chem.* 274(7):4045–52 (1999).

Lewerenz et al., "Shared Receptor Components but Distinct Complexes for α and β Interferons", *J. Mol. Biol.* 282(3):585–99 (1998).

Clemens, *Cytokines*, BIOS Scientific Publishers Limited, Oxford, UK, 1991.

De Maeyer et al., "The Interferon Gene Family" in *Interferons and Other Regulatory Cytokines*, Chap. 2, pp. 5–38, Wiley, New York, 1988.

Tanaguchi et al., "The nucleotide sequence of human fibroblast interferon cDNA", *Gene* 10(1):11–15 (1980).

Houghton et al., "The complete amino acid sequence of human fibroblast interferon as deduced using synthetic oligodeoxyribonucleotide primers of reverse transcriptase", *Nucleic Acids Res.* 8(13):2885–94 (1980).

Ohno and Taniguchi, "Inducer-responsive expression of the cloned human interferon β_1 gene introduced into cultured mouse cells", *Nucleic Acids Res.* 10(3):967–77 (1982).

Smith et al., "Production of Human Beta Interferon in Insect Cells Infected with a Baculovirus Expression Vector", *Mol. Cell. Biol.* 3(12):2156–65 (1983).

Demolder et al., "Human interferon-β, expressed in *Saccharomyces cerevisiae*, is predominantly directed to the vacuoles. Influence of modified co-expression of secretion factors and chaperone", *J. Biotechnol.* 32(2):179–89 (1994).

Desmyter et al., "Administration of Human Fibroblast Interferon in Chronic Hepatitis-B Infection", *Lancet* 2(7987):645–7 (1976).

Makower and Wadler, "Interferons as Biomodulators of Fluoropyrimidines in the Treatment of Colorectal Cancer", *Semin. Oncol.* 26(6):663–71 (1999).

(List continued on next page.)

Primary Examiner—Lorraine Spector
Assistant Examiner—Dong Jiang
(74) *Attorney, Agent, or Firm*—Richard F. Trecartin; Robin M. Silva; Renee M. Kosslrk

(57) **ABSTRACT**

The invention relates to novel interferon-beta activity (IbA) proteins and nucleic acids. The invention further relates to the use of the IbA proteins in the treatment of IFN-β related disorders.

17 Claims, 23 Drawing Sheets

OTHER PUBLICATIONS

Sturzebecher et al., "Pharmacodynamic Comparison of Single Doses on IFN–β1a abd UFB–β1b in Healthy Volunteers", *J. Interferon Cytokine Res.* 19(11):1257–64 (1999).

Zein, "Interferons in the management of viral hepatitis", Cytokines Cell. Mol. Ther. 4(4):229–41 (1998).

Musch et al., "Phase II Clinical Trial of Combined Natural Interferon–β plus Recombinant Interferon–y Treatment of Chronic Hepatitis B", *Hepato–Gastroenterology* 45(24):2282–94 (1998).

Wadler et al., "Sequential Phase II Trials of Fluorouracil and Interferon β_{ser} with or without Sargramostim in Patients with Advanced Colorectal Carcinoma", *Cancer J. Sci. Am.* 4(5):331–7 (1998).

Arnason, "Treatment of multiple sclerosis with interferon β", *Biomed Pharmacother* 53(8):344–50, (1999).

Comi et al., "Interferon beta treatment in multiple sclerosis: the European clinical trials", Mult. Scler. 1(6):317–20 (1996).

Kappos, "Multiple Sclerosis trials", *Lancet* 353(9171):2242–3 (1999).

Senda et al., "Three–dimensional crystal structure of recombinant murine interferon–β", EMBO J. 11(9):3193–3201 (1992).

Senda et al., "Refined Crystal Structure of Recominant Murine Interferon–β at 2.15 Å Resolution", *J. Mol. Biol.* 253(1):187–207 (1995).

Mitsui et al., "Structural, Functional and Evolutionary Implications of the Three–Dimensional Crystal Structure of Murine interfereon–β", *Pharmacol. Ther.* 58(1):93–132 (1993).

Mitsui et al., "Elucidation of the Basic Three–Dimensional Structure of Type I Interferons and Its Functional and Evolutionary Implications", *J. Interferon Cytokine Res.* 17(6):319–26 (1997).

Karpusas et al., "The crystal structure of human interferon β at 2.2–Å resolution", *Proc. Natl. Acad. Sci. U.S.A.* 94(22):11813–8 (1997).

Runkel et al., "Structural and Functional Differences Between Glycosylated and Non–glycosylated Forms of Human Interferon–β (IFN–β)", *Pharm. Res.* 15(4):641–9 (1998).

Runkel et al., "Differences in Activity between α and β Type I Interferons Explored by Mutational Analysis", J. Biol. Chem. 273(14):8003–8 (1998).

Hellinga et al., "Construction of New Ligand Binding Sites in Proteins of Known Structure; I. Computer–aided Modeling of Sites with Pre–defined Geometry", *J. Mol. Biol.* 222:763–785 (1991).

Hurley et al., "Design and Structural Analysis of Alternative Hydrophobic Core Packing Arrangements in Bacteriophage T4 Lysozyme", *J. Mol. Biol.* 224:1143–1154 (1992).

Desjarlais and Handel, "De novo design of the hydrophobic cores of proteins", *Protein Science* 4:2006–2018 (1995).

Harbury et al., "Repacing protein cores with backbone freedom: Structure prediction for coiled coils", *Proc. Natl. Acad. Sci. USA* 92:8408–8412 (1995).

Klemba et al., "Novel metal–binding proteins by design", *Struc. Biol.* 2(5):368–373 (1995).

Nautiyal, et al., "A Designed Heterotrimetic Coiled Coil", *Biochemistry* 34:11645–11651 (1995).

Betz and Grado, "Controlling Topology and Native–like Behavior fo de Novo–Designed Peptides: Design and Characerization of Antiparallel Four–Stranded Coiled Coils", *Biochemistry* 35:6955–6962 (1996).

Dahiyat and Mayo, "Protein Design automation", *Protein Science* 5:895–903 (1996).

Dahiyat and Mayo, "De Novo Protein Design: Fully Automated Sequence Selection", *Science* 278:82–87 (1997).

Dahayat et al., "De Novo Protein Design: Towards Fully Automated Sequence Selection", *J. Mol. Biol.* 273:789–796 (1997).

Dahiyat et al., "Automated design of the surface positions of protein helices", *Protein Science* 6:1333–1337 (1997).

Jones, "De novo protein design using pairwise potentials and a genetic algorith", *Protein Science* 3:567–574 (1994).

Kono and Doi, "Energy Minimization Method Using Automata Network for Sequence and Side–Chain Conformation Prediction From Given Backbone Geometry", *Proteins: Structure, Function and Genetics* 19:244–255 (1994).

Runkel, et al., "Systematic Mutational Mapping of Sites on Human Interferon–β–1a That Are Important for Receptor Binding and Functional Activity." *Biochemistry* 39:2538–2551 (2000).

* cited by examiner

Chain-A: Sequence and Secondary Structure

```
  1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
      HHHHHHHH HHHHHHHHHH HTTS    SG GGGG        HHHHH

 51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETIVENLLAN VYHQINHLKT
     HHHHHHHHH HHHHHHHHHH TS  GGGT   HHHHHHHHHH HHHHHHHHHH

101 VLEEKLEKED FTRGKLMSSL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
    HHHHHHTTSS      SSSHH HHHHHHHHHH HHHHHTTT H HHHHHHHHHH

151 LRNFYFINRL TGYLRN
    HHHHHHHHHH HTT
```

FIG._1A

Chain-B: Sequence and Secondary Structure

```
  1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
      HHHHHHHH HHHHHHHHHH HHH           S        HHHH  S

 51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETIVENLLAN VYHQINHLKT
     HHHHHHHHH HHHHHHHHHH HS  TTT    HHHHHHHHHH HHHHHHHHHH

101 VLEEKLEKED FTRGKLMSSL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
    HHHHTTTTS    HHHHHHHH HHHHHHHHHH HHHHHTTT H HHHHHHHHHH

151 LRNFYFINRL TGYLRN
    HHHHHHHHHH HTT
```

FIG._1B

Human Interferon-Beta Gene Sequence

```
  1 atgaccaaca agtgtctcct ccaaattgct ctcctgttgt gcttctccac tacagctctt
 61 tccatgagct acaacttgct tggattccta caaagaagca gcaattttca gtgtcagaag
121 ctcctgtggc aattgaatgg gaggcttgaa tattgcctca aggacaggat gaactttgac
181 atccctgagg agattaagca gctgcagcag ttccagaagg aggacgccgc attgaccatc
241 tatgagatgc tccagaacat ctttgctatt ttcagacaag attcatctag cactggctgg
301 aatgagacta ttgttgagaa cctcctggct aatgtctatc atcagataaa ccatctgaag
361 acagtcctgg aagaaaaact ggagaaagaa gattttacca ggggaaaact catgagcagt
421 ctgcacctga aaagatatta tgggaggatt ctgcattacc tgaaggccaa ggagtacagt
481 cactgtgcct ggaccatagt cagagtggaa atcctaagga acttttactt cattaacaga
541 cttacaggtt acctccgaaa ctgaagatct cctagcctgt ccctctggga ctggacaatt
601 gcttcaagca ttcttcaacc agcagatgct gtttaagtga ctgatggcta atgtactgca
661 aatgaaagga cactagaaga ttttgaaatt tttattaaat tatgagttat ttttatttat
721 ttaaatttta ttttggaaaa taaattattt ttggtgc
```

FIG._1C

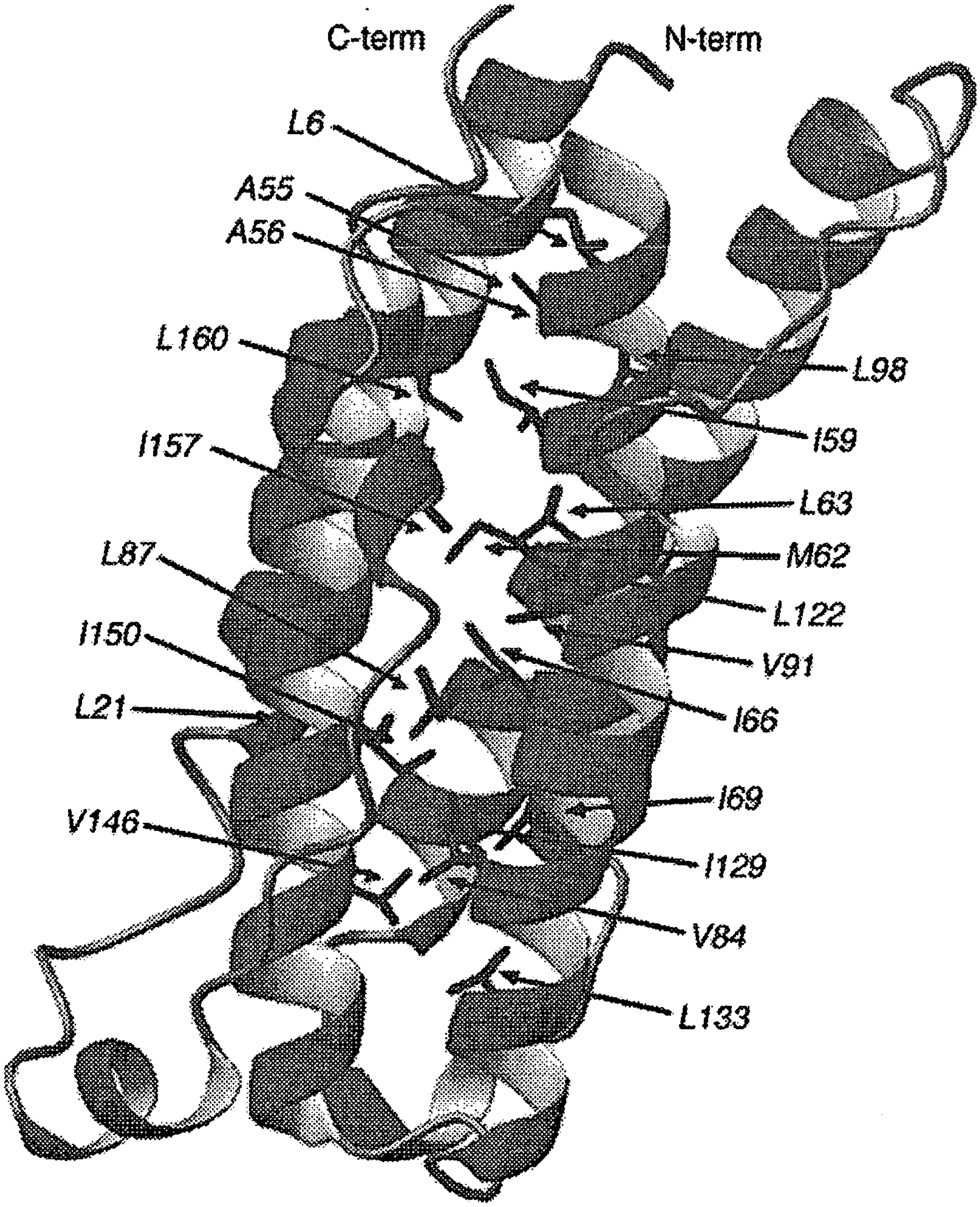

C-term
N-term
L6
A55
A56
L160
I157
L87
I150
L21
V146
L98
I59
L63
M62
L122
V91
I66
I69
I129
V84
L133

FIG._2

IFNβ Core 1 (A-Chain, B-Chain)

```
 6   21  55  56  59  62  63  66  69  84  87  91  98 122 129 133 146 150 157 160
Leu Leu Ala Ala Ile Met Leu Ile Ile Val Leu Val Leu Leu Ile Leu Val Ile Ile Leu
```

IFNβ Core 2 (A-Chain, B-Chain)

```
 1   6   10  14  17  21  38  50  55  56  58  59  61  62  63  66  69  70  81  84
Met Leu Gln Asn Cys Leu Phe Phe Ala Ala Thr Ile Glu Met Leu Ile Ile Phe Glu Val
87   91  94  95  98 102 115 122 125 126 129 130 133 138 144 146 147 150 151 153
Leu Val Gln Ile Leu Leu Lys Leu Tyr Tyr Ile Leu Leu Tyr Thr Val Arg Ile Leu Asn
154 157 159 160 161 163 164
Phe Ile Arg Leu Thr Tyr Leu
```

IFNβ Core 3, Core 4, Core 5, Core 6 (A-Chain); Core 5, Core 6, Core 7 (B-Chain)

```
 1   6   10  13  14  17  18  21  38  50  55  56  58  59  61  62  63  66  69  70
Met Leu Gln Ser Asn Cys Gln Leu Phe Phe Ala Ala Thr Ile Glu Met Leu Ile Ile Phe
72   74  76  77  81  84  87  90  91  94  95  98 102 114 115 118 122 125 126 129
Gln Ser Ser Thr Glu Val Leu Asn Val Gln Ile Leu Leu Gly Lys Ser Leu Tyr Tyr Ile
130 132 133 136 138 139 142 143 144 146 147 150 151 153 154 157 159 160 161 163
Leu Tyr Leu Lys Tyr Ser Ala Trp Thr Val Arg Ile Leu Asn Phe Ile Arg Leu Thr Tyr
164
Leu
```

IFNβ Core 3, Core 4 (B-Chain)

```
 1   6   10  13  14  15  17  21  38  50  55  56  58  59  61  62  63  66  69  70
Met Leu Gln Ser Asn Phe Cys Leu Phe Phe Ala Ala Thr Ile Glu Met Leu Ile Ile Phe
72   74  76  77  81  84  87  90  91  94  95  98 102 114 115 118 122 125 126 129
Gln Ser Ser Thr Glu Val Leu Asn Val Gln Ile Leu Leu Gly Lys Ser Leu Tyr Tyr Ile
130 132 133 136 138 139 142 143 144 146 147 150 151 153 154 157 159 160 161 163
Leu Tyr Leu Lys Tyr Ser Ala Trp Thr Val Arg Ile Leu Asn Phe Ile Arg Leu Thr Tyr
164
Leu
```

FIG._3

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  6 L:85.0 A:14.7 F:  .3
 21 L:98.7 I:  .7 V:  .4 A:  .1 Y:  .1
 55 A:100.0
 56 A:100.0
 59 I:66.1 V:32.3 A:  .9 L:  .7
 62 M:96.3 I: 3.3 V:  .4
 63 L:99.6 A:  .4
 66 I:48.9 L:45.5 V: 5.4 A:  .2
 69 I:77.7 V:19.5 L: 2.8
 84 I:40.5 L:39.4 V:19.6 A:  .5
 87 F:78.7 L:18.4 I: 2.1 Y:  .6 V:  .2
 91 V:83.0 I:16.1 A:  .9
 98 L:97.6 A: 2.4
122 L:100.0
129 I:75.6 V:20.7 L: 2.9 A:  .8
133 L:100.0
146 V:83.8 I:15.5 A:  .7
150 I:81.6 V:18.2 A:  .2
157 I:81.5 V:13.4 L: 4.7 A:  .4
160 L:73.5 I:16.5 V: 9.4 A:  .6
```

FIG._4A

```
  1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETIIENFLAN VYHQINHLKT
101 VLEEKLEKED FTRGKLMSSL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL TGYLRN
```

FIG._4B

```
Res  Cons
Num  Seq   Other Mutations
^^^  ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:99.1 A:  .9
 10 Q:100.0
 14 N:100.0
 17 C:100.0
 21 L:59.3 I:39.7 V:  .9 F:  .1
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 A:100.0
 58 T:100.0
 59 I:93.4 V: 6.4 L:  .2
 61 E:100.0
 62 M:98.4 L: 1.2 I:  .3 V:  .1
 63 L:88.6 F:11.4
 66 I:85.1 L: 9.9 V: 5.0
 69 I:82.7 V:15.7 L: 1.6
 70 F:100.0
 81 E:100.0
 84 I:76.5 L:14.1 V: 8.8 A:  .6
 87 L:69.9 F:11.8 I:10.1 V: 8.0 Y:  .1 A:  .1
 91 I:81.7 V:11.5 L: 6.8
 94 Q:100.0
 95 I:100.0
 98 F:68.8 L:31.2
102 L:100.0
115 K:100.0
122 L:50.7 I:27.4 V:20.9 A: 1.0
125 Y:100.0
126 Y:100.0
129 I:60.7 L:34.8 V: 4.5
130 L:100.0
133 L:100.0
138 Y:100.0
144 T:100.0
146 V:95.2 I: 3.5 A: 1.3
147 R:100.0
150 I:68.3 L:16.5 A:12.2 V: 3.0
151 L:100.0
153 N:100.0
154 F:100.0
157 L:51.4 I:42.7 V: 5.8 A:  .1
159 R:100.0
160 L:86.8 I: 9.9 V: 3.3
161 T:100.0
163 Y:100.0
164 L:100.0
```

FIG._5A

1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETIIENLLAN IYHQINHFKT

101 VLEEKLEKED FTRGKLMSSL HIKRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYFLNRL TGYLRN

FIG._5B

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:97.6 F: 2.4
 10 Q:100.0
 13 F:67.7 Y:31.4 L:  .7 I:  .2
 14 N:100.0
 17 C:88.7 A: 6.9 L: 3.9 V:  .5
 18 Q:100.0
 21 L:85.4 I:14.0 V:  .5 F:  .1
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 A:100.0
 58 T:100.0
 59 I:81.4 V:15.9 L: 2.3 A:  .4
 61 E:100.0
 62 M:91.3 I: 8.7
 63 L:69.8 F:29.8 Y:  .4
 66 I:91.6 V: 7.8 L:  .6
 69 I:66.8 V:33.2
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:98.7 L: 1.3
 87 L:73.8 I:15.1 V:10.4 F:  .7
 90 N:100.0
 91 I:66.4 L:19.2 V:13.8 A:  .3 F:  .3

 94 Q:100.0
 95 I:100.0
 98 L:62.1 F:35.4 A: 2.5
102 L:100.0
114 G:100.0
115 K:100.0
118 A:89.1 V:10.9
122 L:68.8 I:20.3 V:10.9
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 I:100.0
147 R:100.0
150 I:96.0 L: 3.2 A:  .8
151 L:100.0
153 N:100.0
154 F:97.2 L: 1.7 Y: 1.1
157 L:42.1 I:33.9 V:20.8 A: 3.2
159 R:100.0
160 L:86.6 I:13.4
161 A:50.6 V:26.2 I:23.2
163 Y:100.0
164 L:100.0
```

FIG._6A

1 MSYNLLGFLQ RS**F**NFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFA**V**F RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._6B

1 MSYNLLGFLQ RS**Y**NFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFA**V**F RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**V**SL H**V**KRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._6C

1 MSYNLLGFLQ RS**F**NFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._6D

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:98.1 F: 1.9
 10 Q:100.0
 13 F:67.3 Y:32.7
 14 N:100.0
 17 D:82.9 T: 7.1 A: 4.5 L: 4.1 V: 1.4
 18 Q:100.0
 21 L:85.8 I:13.7 V:  .5
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 A:100.0
 58 T:100.0
 59 I:77.8 V:19.1 L: 2.6 A:  .5
 61 E:100.0
 62 M:92.1 I: 7.9
 63 L:73.2 F:25.9 Y:  .9
 66 I:93.3 V: 5.9 L:  .8
 69 I:67.7 V:32.3
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:99.9 L:  .1
 87 L:75.1 I:14.7 V: 8.8 F: 1.4
 90 N:100.0
 91 I:75.0 V:14.6 L:10.3 A:  .1
```

```
 94 Q:100.0
 95 I:100.0
 98 L:65.4 F:32.7 A: 1.9
102 L:100.0
114 G:100.0
115 K:100.0
118 A:89.6 V:10.4
122 L:70.2 I:19.4 V:10.4
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 I:100.0
147 R:100.0
150 I:98.4 A: 1.4 L:  .2
151 L:100.0
153 N:100.0
154 F:96.3 Y: 2.0 L: 1.7
157 L:40.1 I:36.3 V:19.8 A: 3.8
159 R:100.0
160 L:86.2 I:13.8
161 A:48.1 V:22.5 I:22.3 T: 5.7 D: 1.4
163 Y:100.0
164 L:100.0
```

FIG._7A

1 MSYNLLGFLQ RS**F**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFA**V**F RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._7B

1 MSYNLLGFLQ RS**Y**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFA**V**F RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**V**SL H**V**KRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._7C

1 MSYNLLGFLQ RS**F**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINH**F**KT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._7D

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:96.2 F: 3.8
 10 Q:100.0
 13 E:63.4 A:34.0 S: 1.3 G:  .8 T:  3 C: .2
 14 N:100.0
 17 C:55.8 D:38.7 A: 5.5
 18 Q:100.0
 21 L:85.3 I:13.7 V:  .8 A:  .2
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 A:100.0
 58 T:100.0
 59 I:71.6 V:27.9 L:  .5
 61 E:100.0
 62 M:77.4 I:14.4 L: 8.2
 63 F:59.8 L:40.2
 66 I:85.8 V:13.5 L:  .7
 69 I:73.1 V:26.9
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:99.5 L:  .5
 87 L:75.4 I:16.7 V: 7.9
 90 N:100.0
 91 I:63.6 L:25.0 V:11.4
 94 Q:100.0
 95 I:100.0
 98 L:94.5 A: 5.5
102 L:100.0
114 G:100.0
115 K:100.0
118 C:100.0
122 L:77.2 I:22.8
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 I:100.0
147 R:100.0
150 I:98.2 L: 1.5 V:  .3
151 L:100.0
153 N:100.0
154 F:97.9 Y: 1.1 L: 1.0
157 I:43.8 V:41.7 L:13.9 A:  .6
159 R:100.0
160 L:78.6 I:21.4
161 C:99.8 A:  .2
163 Y:100.0
164 L:100.0
```

FIG._8A

1 MSYNLLGFLQ RS**E**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**C**SL HLKRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYFINRL **C**GYLRN

FIG._8B

1 MSYNLLGFLQ RS**A**NFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**C**SL HLKRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **C**GYLRN

FIG._8C

1 MSYNLLGFLQ RS**E**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**C**SL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYFINRL **C**GYLRN

FIG._8D

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:96.8 F: 3.2
 10 Q:100.0
 13 E:66.7 A:31.9 S: 1.2 T:  .1 G:  .1
 14 N:100.0
 17 D:80.8 A:10.1 T: 9.1
 18 Q:100.0
 21 L:88.1 I:11.6 V:  .3
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 A:100.0
 58 T:100.0
 59 I:69.4 V:29.2 L: 1.3 A:  .1
 61 E:100.0
 62 M:85.3 I:14.7
 63 L:52.9 F:47.1
 66 I:90.9 V: 8.6 L:  .5
 69 I:79.1 V:20.9
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:99.6 L:  .4
 87 L:77.8 I:15.1 V: 7.1
 90 N:100.0
 91 I:72.1 L:16.2 V:11.7
```

```
 94 Q:100.0
 95 I:100.0
 98 L:96.6 A: 3.4
102 L:100.0
114 G:100.0
115 K:100.0
118 A:100.0
122 L:84.2 I:15.8
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 I:100.0
147 R:100.0
150 I:99.1 L:  .9
151 L:100.0
153 N:100.0
154 F:98.5 Y: 1.1 L:  .4
157 I:43.3 V:36.3 L:19.5 A:  .9
159 R:100.0
160 L:79.6 I:20.4
161 A:55.1 T:23.9 D:21.0
163 Y:100.0
164 L:100.0
```

FIG._9A

1 MSYNLLGFLQ RS**E**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFA**V**F RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._9B

1 MSYNLLGFLQ RS**E**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**A**SL HLKRYYGRIL HYLKAKEYSH CAWTI**I**RVEI

151 LRNFYF**L**NRL TGYLRN

FIG._9C

1 MSYNLLGFLQ RS**E**NFQ**D**QKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETI**I**ENLLAN **I**YHQINHLKT

101 VLEEKLEKED FTRGKLM**A**SL H**I**KRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYF**L**NRL **A**GYLRN

FIG._9D

```
Res  Cons
Num  Seq   Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  6 L:98.4 A: 1.6
 21 L:100.0
 55 A:100.0
 56 A:100.0
 59 I:78.0 V:21.1 A:   .6 L:   .3
 62 M:84.7 I:14.4 V:   .9
 63 L:84.3 F:15.7
 66 I:53.1 L:42.4 V: 4.5
 69 I:91.2 V: 8.8
 84 I:62.3 V:25.4 L:11.7 A:   .6
 87 F:74.6 L:21.5 W: 1.9 Y: 1.3 I:   .6 V:   .1
 91 I:54.7 V:43.6 L: 1.5 A:   .2
 98 L:98.1 A: 1.9
122 L:82.8 F:13.6 I: 3.6
129 I:77.5 V:22.5
133 L:100.0
146 V:99.7 A:   .3
150 I:88.5 V:11.0 L:   .5
157 I:78.4 V:15.1 L: 6.5
160 L:59.2 F:39.4 Y: 1.3 A:   .1
```

FIG._10A

```
  1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF

 51 QKEDAALTIY EMLQNIFAIF RQDSSSTGWN ETIIENFLAN VYHQINHLKT

101 VLEEKLEKED FTRGKLMSSL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI

151 LRNFYFINRL TGYLRN
```

FIG._10B

```
Res  Cons
Num  Seq  Other Mutations                        94 Q:100.0
^^^ ^^^^^ ^^^^^^^^^^^^^^^                        95 I:100.0
  1 M:100.0                                      98 L:97.4 F: 1.7 A:  .9
  6 L:98.5 A: 1.5                               102 L:100.0
 10 Q:100.0                                     115 K:100.0
 14 N:100.0                                     122 F:35.3 I:28.3 L:26.2 Y: 6.9 V: 2.7 W:  .6
 17 C:100.0                                     125 Y:100.0
 21 L:84.6 F:15.4                               126 Y:100.0
 38 F:100.0                                     129 I:87.5 L: 6.4 V: 6.1
 50 F:100.0                                     130 L:100.0
 55 A:100.0                                     133 L:100.0
 56 L:97.6 A: 2.4                               138 Y:100.0
 58 T:100.0                                     144 T:100.0
 59 I:89.3 V: 8.6 A: 2.1                        146 V:97.6 I: 1.6 A:  .8
 61 E:100.0                                     147 R:100.0
 62 M:84.6 L:11.1 I: 3.4 V:  .9                 150 I:95.7 V: 3.5 L:  .8
 63 L:67.2 F:32.4 Y:  .4                        151 L:100.0
 66 I:93.1 L: 3.6 V: 3.3                        153 N:100.0
 69 I:90.4 V: 9.6                               154 F:100.0
 70 F:100.0                                     157 I:88.1 L: 7.4 V: 4.5
 81 E:100.0                                     159 R:100.0
 84 I:74.9 V:15.5 L: 8.4 A: 1.2                 160 L:65.0 F:31.5 Y: 1.9 I: 1.4 A:  .2
 87 F:69.3 L:24.4 I: 5.5 V:  .4 Y:  .4          161 T:100.0
 91 I:68.5 L:27.7 V: 3.8                        163 Y:100.0
                                                164 L:100.0
```

FIG._11A

```
  1 MSYNLLGFLQ RSSNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMFQNIFAIF RQDSSSTGWN ETIIENFLAN IYHQINHLKT
101 VLEEKLEKED FTRGKLMSSL HFKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL TGYLRN
```

FIG._11B

```
Res  Cons
Num  Seq   Other Mutations
^^^  ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:100.0
 10 Q:100.0
 13 L:92.7 A: 7.3
 14 N:100.0
 15 F:100.0
 17 C:64.8 A:25.8 V: 6.1 L: 2.2 I: 1.1
 21 L:85.8 F:14.1 Y:  .1
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 L:99.8 A:  .2
 58 T:100.0
 59 I:98.0 A: 1.9 L:  .1
 61 E:100.0
 62 M:79.9 I:12.2 L: 7.4 V:  .5
 63 L:75.4 F:22.9 Y: 1.7
 66 I:73.8 L:15.4 V:10.8
 69 I:96.7 A: 1.7 L: 1.6
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:100.0
 87 F:46.8 L:45.0 I: 7.1 V:  .6 Y:  .4 W:  .1
 90 N:100.0
 91 I:52.6 L:27.8 V:15.1 F: 4.3 Y:  .1 A:  .1
```

```
 94 Q:100.0
 95 I:100.0
 98 L:97.8 F: 2.2
102 L:100.0
114 F:100.0
115 K:100.0
118 L:100.0
122 I:39.9 L:39.0 F:21.1
125 Y:100.0
126 Y:100.0
129 I:99.1 L:  .9
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 V:99.1 I:  .9
147 R:100.0
150 I:71.6 L:14.2 V:12.5 F: 1.7
151 L:100.0
153 N:100.0
154 F:89.9 L: 8.9 Y: 1.2
157 I:62.2 L:28.0 V: 9.7 A:  .1
159 R:100.0
160 L:97.3 F: 2.7
161 A:100.0
163 Y:100.0
164 L:100.0
```

FIG._12A

```
  1 MSYNLLGFLQ RSLNFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMLQNIFAIF RQDSSSTGWN ETIIENLLAN IYHQINHLKT
101 VLEEKLEKED FTRFKLMLSL HIKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL AGYLRN
```

FIG._12B

```
Res  Cons
Num  Seq   Other Mutations
^^^  ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:98.6 F: 1.4
 10 Q:100.0
 13 L:42.5 E:25.4 S:25.1 T:3.1 D: 2.4 A: 1.5
 14 N:100.0
 15 F:100.0
 17 A:53.0 T:24.0 V: 7.6 I:6.5 L: 5.7 D: 3.1 E:  .1
 21 L:94.2 F: 5.8
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 L:97.7 A: 2.3
 58 T:100.0
 59 I:92.0 V: 8.0
 61 E:100.0
 62 M:82.1 L:11.6 I: 6.0 V:  .3
 63 L:75.3 F:24.7
 66 I:85.2 L:12.2 V: 2.6
 69 I:100.0
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:100.0
 87 F:80.7 L:14.1 I: 4.8 Y:  .4
 90 N:100.0
 91 I:53.7 L:35.3 V: 9.8 A:  .9 F:  .3
```

```
 94 Q:100.0
 95 I:100.0
 98 L:97.9 A: 2.1
102 L:100.0
114 F:100.0
115 K:100.0
118 L:100.0
122 I:46.0 L:30.0 F:24.0
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 V:100.0
147 R:100.0
150 I:92.3 V: 5.0 L: 2.7
151 L:100.0
153 N:100.0
154 F:88.6 L:11.4
157 I:72.3 L:21.2 V: 6.5
159 R:100.0
160 L:74.2 F:25.8
161 E:85.0 T:15.0
163 Y:100.0
164 L:100.0
```

FIG._13A

```
  1 MSYNLLGFLQ RSLNFQAQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMLQNIFAIF RQDSSSTGWN ETIIENFLAN LYHQINHLKT
101 VLEEKLEKED FTRFKLMLSL HIKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL EGYLRN
```

FIG._13B

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:99.0 F:  .8 A:  .2
 10 Q:100.0
 13 E:38.9 C:36.6 S:21.7 D: 2.8
 14 N:100.0
 17 C:91.2 A: 5.1 D: 3.1 T:  .6
 18 Q:100.0
 21 L:72.1 F:27.9
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 L:97.6 A: 2.4
 58 T:100.0
 59 I:98.5 V: 1.5
 61 E:100.0
 62 M:81.8 L:10.2 I: 8.0
 63 L:83.9 F:16.1
 66 I:89.6 L: 8.0 V: 2.4
 69 I:99.6 A:  .2 L:  .2
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:82.4 V:13.8 L: 3.8
 87 L:93.9 I: 6.0 V:  .1
 90 N:100.0
 91 I:58.7 L:19.0 F:17.2 V: 5.1
```

```
 94 Q:100.0
 95 I:100.0
 98 L:98.9 A: 1.1
102 L:100.0
114 L:100.0
115 K:100.0
118 E:92.8 C: 7.2
122 L:40.7 I:31.4 F:27.1 W:  .8
125 Y:100.0
126 Y:100.0
129 I:99.8 L:  .2
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 V:99.8 I:  .2
147 R:100.0
150 I:91.3 L: 8.3 F:  .2 V:  .2
151 L:100.0
153 N:100.0
154 F:82.7 L:17.3
157 I:69.1 L:24.5 V: 6.4
159 R:100.0
160 L:87.6 F:10.9 I: 1.5
161 E:84.3 T:15.7
163 Y:100.0
164 L:100.0
```

FIG._14A

```
  1 MSYNLLGFLQ RSENFQCQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMLQNIFAIF RQDSSSTGWN ETIIENLLAN IYHQINHLKT
101 VLEEKLEKED FTRLKLMESL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL EGYLRN
```

FIG._14B

```
Res  Cons
Num  Seq  Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:98.7 F: 1.3
 10 Q:100.0
 13 S:49.4 E:33.2 T: 7.9 D: 5.2 A: 4.3
 14 N:100.0
 17 T:36.3 D:29.4 A:29.3 E: 4.3 S:  .7
 18 Q:100.0
 21 L:78.3 F:21.6 Y:  .1
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 L:98.1 A: 1.9
 58 T:100.0
 59 I:98.6 V: 1.4
 61 E:100.0
 62 M:82.4 L:12.1 I: 5.5
 63 L:78.7 F:21.3
 66 I:90.4 L: 6.0 V: 3.6
 69 I:100.0
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:94.0 L: 6.0
 87 L:93.4 I: 6.6
 90 N:100.0
 91 I:76.4 L:11.7 F: 8.1 V: 3.8
 94 Q:100.0
 95 I:100.0
 98 L:97.9 A: 2.1
102 L:100.0
114 L:100.0
115 K:100.0
118 E:99.4 A:  .6
122 L:41.3 I:38.9 F:19.2 W:  .6
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 V:100.0
147 R:100.0
150 I:83.4 L:15.6 V: 1.0
151 L:100.0
153 N:100.0
154 F:87.2 L:12.6 Y:  .2
157 I:65.6 L:27.6 V: 6.8
159 R:100.0
160 L:89.6 F:10.4
161 E:86.4 T:12.1 G: 1.5
163 Y:100.0
164 L:100.0
```

FIG._15A

```
  1 MSYNLLGFLQ RSSNFQTQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMLQNIFAIF RQDSSSTGWN ETIIENLLAN IYHQINHLKT
101 VLEEKLEKED FTRLKLMESL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL EGYLRN
```

FIG._15B

FIG._15B

```
Res  Cons
Num  Seq   Other Mutations
^^^ ^^^^^ ^^^^^^^^^^^^^^^
  1 M:100.0
  6 L:96.9 F: 3.1
 10 Q:100.0
 13 S:47.4 E:35.2 T: 7.7 D: 6.1 A: 3.6
 14 N:100.0
 17 T:32.8 A:31.0 D:29.0 E: 5.0 S: 1.4 G:  .8
 18 Q:100.0
 21 L:77.9 F:22.0 Y:  .1
 38 F:100.0
 50 F:100.0
 55 A:100.0
 56 L:97.6 A: 2.4
 58 T:100.0
 59 I:99.9 V:  .1
 61 E:100.0
 62 M:78.5 L:13.5 I: 8.0
 63 L:80.7 F:19.3
 66 I:85.6 L: 7.8 V: 6.6
 69 I:98.8 A: 1.2
 70 F:100.0
 72 Q:100.0
 74 S:100.0
 76 S:100.0
 77 T:100.0
 81 E:100.0
 84 I:99.7 L:  .3
 87 L:92.7 I: 7.3
 90 N:100.0
 91 I:73.3 L:13.3 F: 8.7 V: 4.7

 94 Q:100.0
 95 I:100.0
 98 L:96.4 A: 3.6
102 L:100.0
114 G:100.0
115 K:100.0
118 E:100.0
122 L:43.6 I:38.0 F:18.4
125 Y:100.0
126 Y:100.0
129 I:100.0
130 L:100.0
132 Y:100.0
133 L:100.0
136 K:100.0
138 Y:100.0
139 S:100.0
142 A:100.0
143 W:100.0
144 T:100.0
146 V:100.0
147 R:100.0
150 I:76.2 L:17.8 V: 5.4 F:  .6
151 L:100.0
153 N:100.0
154 F:85.5 L:14.1 Y:  .4
157 I:65.7 L:26.6 V: 7.7
159 R:100.0
160 L:95.2 I: 4.8
161 E:98.1 G: 1.9
163 Y:100.0
164 L:100.0
```

FIG._16A

```
  1 MSYNLLGFLQ RSSNFQTQKL LWQLNGRLEY CLKDRMNFDI PEEIKQLQQF
 51 QKEDALLTIY EMLQNIFAIF RQDSSSTGWN ETIIENLLAN IYHQINHLKT
101 VLEEKLEKED FTRGKLMESL HLKRYYGRIL HYLKAKEYSH CAWTIVRVEI
151 LRNFYFINRL EGYLRN
```

FIG._16B

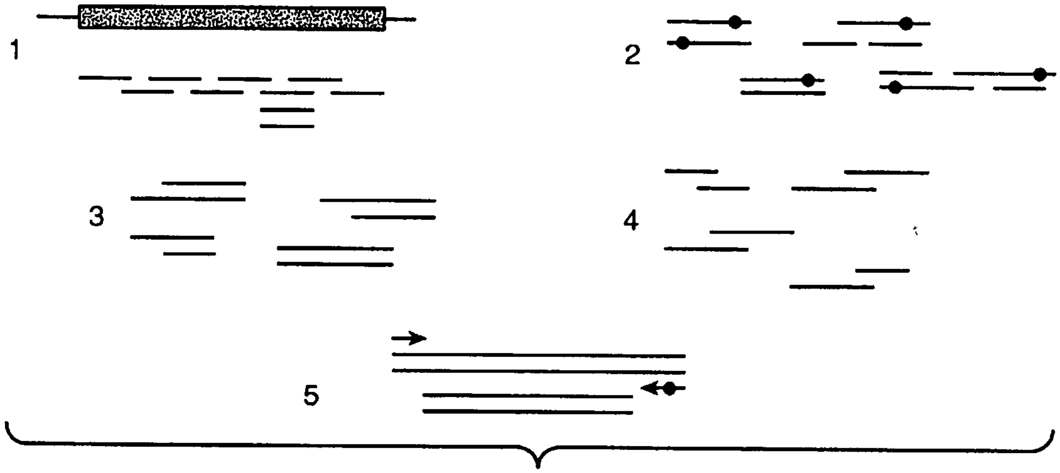

1
2
3
4
5

FIG._17

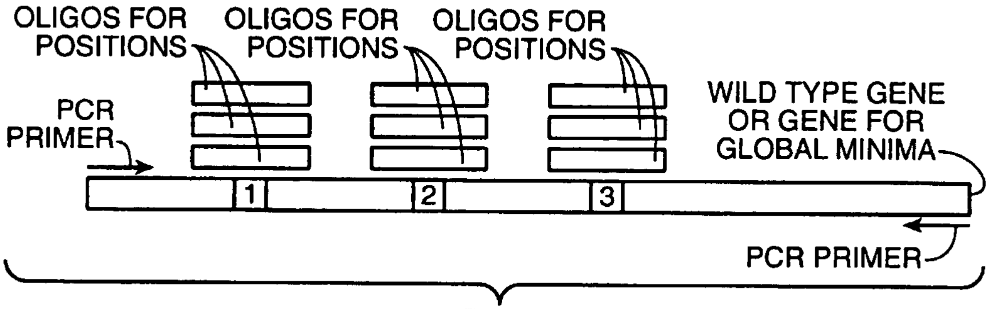
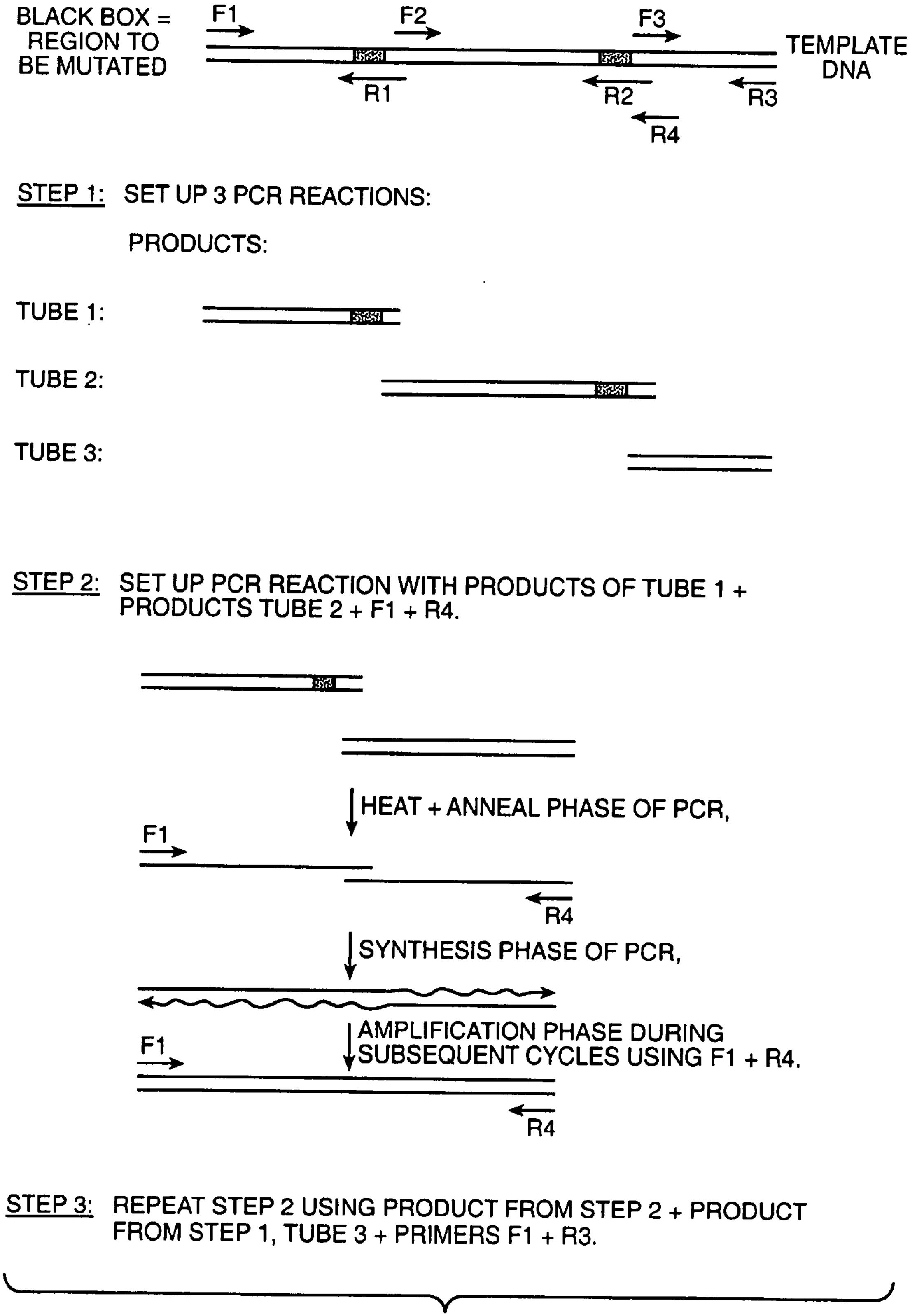

OLIGOS FOR POSITIONS
OLIGOS FOR POSITIONS
OLIGOS FOR POSITIONS
PCR PRIMER
WILD TYPE GENE OR GENE FOR GLOBAL MINIMA
1
2
3
PCR PRIMER

FIG._18



STEP 1: SET UP 3 PCR REACTIONS:

STEP 2: SET UP PCR REACTION WITH PRODUCTS OF TUBE 1 + PRODUCTS TUBE 2 + F1 + R4.

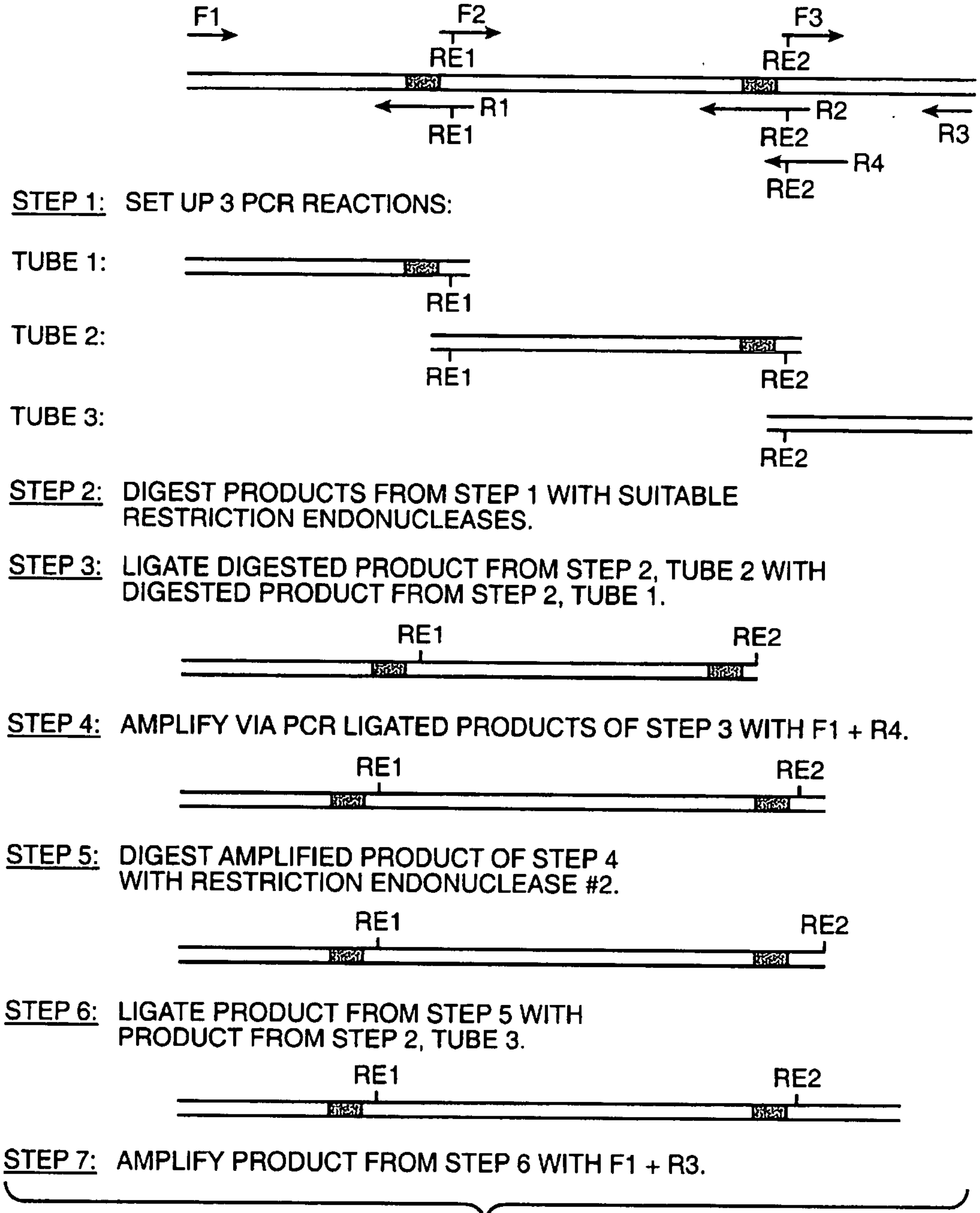
HEAT + ANNEAL PHASE OF PCR,

SYNTHESIS PHASE OF PCR,

AMPLIFICATION PHASE DURING SUBSEQUENT CYCLES USING F1 + R4.

STEP 3: REPEAT STEP 2 USING PRODUCT FROM STEP 2 + PRODUCT FROM STEP 1, TUBE 3 + PRIMERS F1 + R3.

FIG._19



STEP 1: SET UP 3 PCR REACTIONS:

STEP 2: DIGEST PRODUCTS FROM STEP 1 WITH SUITABLE RESTRICTION ENDONUCLEASES.

STEP 3: LIGATE DIGESTED PRODUCT FROM STEP 2, TUBE 2 WITH DIGESTED PRODUCT FROM STEP 2, TUBE 1.

STEP 4: AMPLIFY VIA PCR LIGATED PRODUCTS OF STEP 3 WITH F1 + R4.

STEP 5: DIGEST AMPLIFIED PRODUCT OF STEP 4 WITH RESTRICTION ENDONUCLEASE #2.

STEP 6: LIGATE PRODUCT FROM STEP 5 WITH PRODUCT FROM STEP 2, TUBE 3.

STEP 7: AMPLIFY PRODUCT FROM STEP 6 WITH F1 + R3.

FIG._20

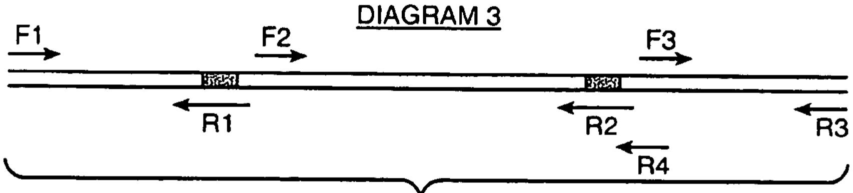


FIG._21

RECOMBINANT INTERFERON-BETA MUTEINS

This application is a continuing application of U.S. Ser. No. 60/133,785, filed May 12, 1999.

FIELD OF THE INVENTION

The invention relates to novel interferon-beta activity (IbA) proteins and nucleic acids. The invention further relates to the use of the IbA proteins in the treatment of interferon-beta (INF-β) related disorders.

BACKGROUND OF THE INVENTION

Human Interferons (IFNs) are members of a biologically potent family of cytokines. Originally, IFNs were identified as agents produced and secreted by virus-infected cells which can protect cells against further viral infections. However, in addition to this antiviral effect, IFNs can elicit many other changes in cellular behavior, including effects on cellular growth and differentiation and modulation of the immune system [e.g., see Lengyel, Annu. Rev. Biochem. 51:251–82 (1982); Gresser and Tovey, Biochim. Biophys. Acta 516(2):231–47 (1978); Gresser et al., Nature New Biol. 231(18):20–1 (1971); Dolei et al., J. Gen. Virol. 46(1):227–36 (1980); Gresser, Cell Immunol 34(2):406–15 (1977)]. By virtue of their antigenic, biological and physicochemical properties, IFNs are classified into three groups, INF-α (leukocyte), INF-β (fibroblast) and INF-γ (immune) [Stewart, J. Infect. Dis. 142(4):643 (1980)].

In humans, the IFN-α subtype encompass a multigene family of about 20 genes, encoding proteins of 166–172 amino acids that are all closely related. In contrast to this diversity, there is only one human interferon-beta (IFN-β) gene, also encoding a protein of 166 amino acids. IFN-β has low homology to the IFN-α family and is an N-linked glycoprotein [Knight, Proc. Natl. Acad. Sci. U.S.A. 73(2):520–523 (1976)]. There is also only one human IFN-γ gene that encodes a polypeptde of 143 amino acids that is glycosylated and forms a dimer in its native state. IFN-γ shows only slight structural similarities to IFN-α or to IFN-β.

All IFN-α and IFN-β (also commonly referred to as type I interferon family) appear to bind to a common high affinity cell surface receptor, a 130 kD glycoprotein that is widely distributed on different cell types and that is distinct from the one bound by IFN-γ. Type-I interferons are recognized by a complex containing the receptor subunits ifnar1 and ifnar2 and their associated Janus tyrosine kinases, Tyk2 and Jak1, that activate the transcription factors STAT1 and STAT2, leading to the formation of the transcription factor complex ISGF3 [interferon-stimulated gene factor 3; Li et al., Biochemie 80(8–9):703–20 (1998); Nadeau et al., J. Biol. Chem. 274(7):4045–52 (1999)]. Three distinct modes of IFN/receptor complex interaction are known: (i) INF-α with ifnar1 and ifnar2; (ii) IFN-β with ifnar1 and ifnar2; and (iii) IFN-β with ifnar2 alone [Lewerenz et al., J. Mol. Biol. 282(3):585–99(1998)]. While Lewerenz et al. suggest that INF-α and IFN-β interact with their receptors in different ways and as such may also signal differently, the events responsible for biological activity beyond receptor binding are poorly understood.

As might be predicted for such a large family of cytokines with almost ubiquitously distributed receptors, IFNs display varied physiological roles. Production of IFN-α or IFN-β is induced by infection, including viral infection or the presence of foreign cell types and antigens. It is not clear what specific molecules are responsible for induction, but double-stranded RNA and cytokines can be good inducers. There is much overlap between different cell types in both the inducers and the species of IFN that is induced. The major cell types that produce IFNs are: lymphocytes, monocytes and macrophages (for IFN-α); fibroblasts and some epithelial cells and lymphoblastoid cells (for IFN-β); and activated T lymphocytes (for IFN-γ).

In addition to the 'classical' anti-viral activities that all IFNs elicit in their target cells, the biological consequences of IFN binding to its receptor can include inhibition of cell proliferation, induction of cell differentiation, changes in cell morphology, enhancement of histocompatibility antigen expression on many cells and stimulation of immunoglobulin-Fc receptor expression on macrophages. B lymphocytes can be induced to increase antibody production by low concentration of IFN-α or IFN-β. An additional effect of IFN-α and IFN-β is activation of natural killer cells that may be responsible for the destruction of virus-infected cells or tumor cells in vivo. Overall, IFNs seem to be of great importance as part of the body's defense against foreign organisms, foreign antigens and abnormal cell types (Clemens, in Cytokines, BIOS Scientific Publishers Limited, 1991; De Maeyer et al., in Interferons and Other Regulatory Cytokines, Wiley, New York, 1988).

INF-α and IFN-β were among the first of the cytokines to be produced by recombinant DNA technology. For example, the amino acid and nucleotide sequence of human IFN-β [Tanaguchi et al., Gene 10(1):11–15 (1980); Houghton et al., Nucleic Acids Res. 8(13):2885–94 (1980)] made it possible to produce recombinant human IFN-β in e.g., mammalian, insect, and yeast cells and in *E. coli*, that is free from viruses and other contaminants from human sources [e.g., Ohno and Taniguchi, Nucleic Acids Res. 10(3):967–77 (1982); Smith et al., Mol. Cell. Biol. 3(12):2156–65 (1983); Demolder et al., J. Biotechnol. 32(2):179–89 (1994); Dorin et al., U.S. Pat. No. 5,814,485 (1998); Konrad et al., U.S. Pat. No. 4,450,103 (1984)].

IFNs have been shown to have therapeutic value in conditions such as inflammatory, viral, and malignant diseases [e.g., see Desmyter et al., Lancet 2(7987):645–7 (1976); Makower and Wadler, Semin. Oncol. 26(6):663–71 (1999); Sturzebecher et al., J. Interferon Cytokine Res. 19(11):1257–64 (1999); Zein, Cytokines Cell. Mol. Ther. 4(4):229–41 (1998; Musch et al., Hepatogastroeneterology 45(24):2282–94 (1998); Wadler et al., Cancer J. Sci. Am. 4(5):331–7 (1998)]. IFN-β is a marketed drug (Betaseron, manufactured by Berlex and Avonex, manufactured by Biogen) that has been approved for use in treatment of multiple sclerosis (MS) [Arnason, Biomed Pharmacother 53(8):344–50, (1999); Comi et al., Mult. Scler. 1(6):317–20 (1996); Aappos, Lancet 353(9171):2242–3 (1999)]. IFN-β seems to reduce the number of attacks suffered by patients with relapsing and remitting MS. Betaseron, a recombinant IFN-β expressed in *E. coli*, consists of 165 amino acids (missing the initial methionine) and is genetically engineered so that it contains a serine at position 17, to replace a cysteine. It is a nonglycosylated form of IFN-β. Avonex is a human IFN-β, consisting of 166 amino acids that is produced by recombinant DNA techniques in CHO cells. This is a glycosylated form of IFN-β. Also, recent studies show promising IFN efficacy in treating certain viral diseases, such as Hepatitis B or C, and cancer.

Most cytokines, including IFN-β, have relatively short circulation half-lives since they are produced in vivo to act locally and transiently. To use IFN-β as an effective systemic therapeutic, one needs relatively large doses and frequent administrations. Frequent parenteral administrations are inconvenient and painful. Further, toxic side effects are associated with IFN-β administration which are so severe that some multiple sclerosis patents cannot tolerate the treatment. These side effects are probably associated with administration of a high dosage. In clinical studies it has been found that some patients produce antibodies to IFN-β, which neutralize its biological activity.

Furthermore, it has been observed that dimers and oligomers of microbially produced IFN-β are formed in *E. coli*, rendering purification and separation of IFN-β laborious and time consuming. It also necessitates several additional steps in purification and isolation procedures such as reducing the protein during purification and reoxidizing it to restore it to its original conformation, thereby increasing the possibility of incorrect disulfide bond formation. In addition, and most likely attributable to the above-listed shortcomings, microbially produced recombinant human IFN-β has also been found to exhibit consistently low specific activity. It would be desirable, therefore, to microbially produce a biologically active IFN-β protein that has a reduced or eliminated ability to form intermolecular crosslinks or intramolecular bonds that cause the protein to adopt an undesirable structure.

To this end, variants of IFN-β sequences, applications and production procedures are known; see for example U.S. Pat. Nos. 4,450,103; 4,518,584; 4,588,585; 4,737,462; 4,738,844; 4,738,845; 4,753,795; 4,769,233; 4,793,995; 4,914,033; 4,959,314; 5,183,746; 5,376,567; 5,545,723; 5,730,969; 5,814,485; 5,869,603 and references cited therein.

Recently, the crystal structures of recombinant murine INFβ [Senda et al., EMBO J. 11(9):3193–201 (1992); Mitsui et al., Pharmacol. Ther. 58(1):93–132 (1993); Senda et al., J. Mol. Biol. 253(1):187–207 (1995); Mitsui et al., J. Interferon Cytokine Res. 17(6):319–26 (1997); all of which are expressly incorporated by reference] and human INFβ [Karpusas et al., Proc. Natl. Acad. Sci. U.S.A. 94(22):11813–8 (1997); Runkel et al., Pharm. Res. 15(4):641–9 (1998); Runkel et a 273(14):8003–8 (1998); Lewerenz et al., J. Mol. Biol. 282(3):585–99 (1998); all of which are expressly incorporated by reference] have been solved. Karpusas et al. determined the crystal structure of glycosylated human IFN-β at 2.2 Angstrom resolution by molecular replacement. The molecule adopts a fold similar to that of the previously determined structures of murine IFN-β and human IFN-α2b, but displays several, distinct structural features. Like human IFN-α2b, INF-P contains a zinc-binding site at the interface of the two molecules in the asymmetric unit, however, unlike human IFN-α2b, IFN-β dimerizes with contact surfaces from opposite sides of the molecule. Runkel et al. reported structural and functional differences between glycosylated (IFN,β-1a) and non-glycosylated (IFN-β1b) forms of human IFN-β and suggested that the greater biological activity of INF-β-1a is due to the stabilizing effect of the carbohydrate moiety.

The available crystal structure of INFβ allows further protein design and the generation of more stable proteins or protein variants with an altered activity. Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms (Hellinga et al., J. Mol. Biol. 222: 763–785 (1991); Hurley et al., J. Mol. Biol. 224:1143–1154 (1992); Desjarlaisl et al., Protein Science 4:2006–2018 (1995); Harbury et al., Proc. Natl. Acad. Sci. U.S.A. 92:8408–8412 (1995); Klemba et al., Nat. Struc. Biol. 2:368–373 (1995); Nautiyal et al., Biochemistry 34:11645–11651 (1995); Betzo et al., Biochemistry 35:6955–6962 (1996); Dahiyat et al., Protein Science 5:895–903 (1996); Dahiyat et al., Science 278:82–87 (1997); Dahiyat et al., J. Mol. Biol. 273:789–96; Dahiyat et al., Protein Sci. 6:1333–1337 (1997); Jones, Protein Science 3:567–574 (1994); Konoi, et al., Proteins: Structure, Function and Genetics 19:244–255 (1994)). These algorithms consider the spatial positioning and steric complementarity of side chains by explicitly modeling the atoms of sequences under consideration. In particular, W098/47089, and U.S. Ser. No. 09/127,926 describe a system for protein design; both are expressly incorporated by reference.

A need still exists for proteins exhibiting both significant stability and interferon-beta activity. Accordingly, it is an object of the invention to provide interferon-beta activity (IbA) proteins, nucleic acids and antibodies for the treatment of multiple sclerosis, cancer and viral infections.

SUMMARY OF THE INVENTION

In accordance with the objects outlined above, the present invention provides non-naturally occurring interferon-beta activity (IbA) proteins (e.g. the proteins are not found in nature) comprising amino acid sequences that are less than about 97% identical to human INF-β. The IbA proteins have at least one altered biological property of an INF-β protein; for example, the IbA proteins will be more stable than IFN-β and bind to cells comprising an interferon receptor complex. Thus, the invention provides IbA proteins with amino acid sequences that have at least about 3–5 amino acid substitutions as compared to the INF-β sequence shown in FIG. 1 (SEQ ID NO:1).

In a further aspect, the present invention provides non-naturally occurring IbA conformers that have three dimensional backbone structures that substantially correspond to the three dimensional backbone structure of INFβ. In one aspect, the three dimensional backbone structure of the IbA conformer corresponds substantially to the three dimensional backbone structure of the A-chain of INFβ. In another aspect, the three dimensional backbone structure of the IbA conformer corresponds substantially to the three dimensional backbone structure of the B-chain of INF-β. The amino acid sequence of the IbA conformer and the amino acid sequence of INF-β are less than about 97% identical. In one aspect, at least about 90% of the non-identical amino acids are in a core region of the conformer. In other aspects, the conformer have at least about 100% of the non-identical amino acids are in a core region of the conformer.

In an additional aspect, the changes are selected from the amino acid residues at positions selected from positions 6, 13, 17;21, 56, 59, 61, 62, 63, 66, 69, 84, 87, 91, 98, 102, 114, 118, 122, 129, 146, 150, 154, 157, 160, and 161. In a preferred aspect, the changes are selected from the amino acid residues at positions selected from positions 13, 17, 56, 59, 63, 66, 69, 84, 87, 91, 98, 114, 118, 122, 146, 157, and 161. In one aspect, the changes are selected from the amino acid residues at positions selected from positions 13, 17, 69, 84, 87, 91, 98, 118, 122, 146, 157, and 161. In another aspect, the changes are selected from the amino acid residues at positions selected from positions 13, 17, 56, 84, 87, 91, 114, 118, 122, and 161. Preferred embodiments include at least about 3–5 variations.

In a further aspect, the invention provides recombinant nucleic acids encoding the non-naturally occurring IbA proteins, expression vectors comprising the recombinant nucleic acids, and host cells comprising the recombinant nucleic acids and expression vectors.

In an additional aspect, the invention provides methods of producing the IbA proteins of the invention comprising culturing host cells comprising the recombinant nucleic acids under conditions suitable for expression of the nucleic acids. The proteins may optionally be recovered. In a further aspect, the invention provides pharmaceutical compositions comprising an IbA protein of the invention and a pharmaceutical carrier.

In an additional aspect, the invention provides methods for treating an INFβ responsive condition comprising administering an IbA protein of the invention to a patient. The INFβ condition includes multiple sclerosis, viral infection, or cancer.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A (SEQ ID NO:1) depicts the amino acid sequence of the A-chain of human INFβ as used in the determination of the crystal structure [PDB and GenBank # 1AU1; Karpusas et al., Proc. Natl. Acad. Sci. U.S.A. 94(22):11813–8 (1997)] and secondary structure elements. Secondary structure element legend: H, alpha helix (4-helix); B, residue in isolated beta bridge; E, extended strand, participates in beta ladder; G, 310 helix (3-helix); I, pi helix (5-helix); T, hydrogen bonded turn; S, bend.

FIG. 1B (SEQ ID NO:1) depicts the amino acid sequences of the B-chain of human INF-β as used in the determination of the crystal structure (Karpusas et al., supra) and secondary structure elements.

FIG. 1C (SEQ ID NO:2) depicts the complete DNA sequence encoding wild type human INF-β (GenBank accession number NM_002176). The encoded sequence consists of the signaling sequence, MTNKCLLQIALLLCFSTTALS (SEQ ID NO:3), and the 166 amino acids that constitute the actual protein (see FIGS. 1A and 1B) (SEQ ID NO:1). The DNA sequence of 757 nucleotides includes this coding sequence and a non-translated region. Bases 1 to 63 encode the signaling sequence; bases 64 to 561 encode the actual IFN-β; bases 562 to 564 (TGA) are stop codon; and the rest is untranslated sequence.

FIG. 2 depicts the structure of wild type IFN-β. Presented is the A-chain from the PDB file 1AU1. The amino acid side chains indicated are those positions included in the PDA design of CORE 1.

FIG. 3 depicts the residues for both the A-chain and B-chain of INF-β selected for PDA. The individual sets are described in detail herein.

FIG. 4A depicts the mutation pattern of IFN-β A-chain core 1 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 1 sequences (only the amino acid residues of positions 6, 21, 55, 56, 59, 62, 63, 66, 69, 84, 87, 91, 98, 122, 129, 133, 146, 150, 157, and 160 are given). All values are given in %. For example, at position 87, the human INF-β amino acid is leucine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 78.7% of the top 1000 sequences had phenylalanine at this position, and only 18.4% of the sequences had leucine. Similarly, for position 84 (valine in human INFβ, isoleucine (40.5%) and leucine (39.4%) are preferred over valine (19.6%).

FIG. 4B (SEQ ID NO:4) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 1 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 5A depicts the mutation pattern of IFN-β A-chain core 2 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 2 sequences (only the amino acid residues of positions 1, 6,10, 14, 17, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 81, 84, 87, 91, 94, 95, 98, 102, 115, 122, 125, 126, 129, 130, 133, 138, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 are given). All values are given in %. For example, at position 91, the human IFN-β amino acid is valine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 81.7% of the top 1000 sequences had isoleucine at this position, and only 11.5% of the sequences had valine. Similarly, for position 98 (leucine in human IFN-β), phenylalanine (68.8%) is preferred over leucine (31.2%).

FIG. 5B (SEQ ID NO:5) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 2 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 6A depicts the mutation pattern of IFN-β A-chain core 3 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 3 sequences (only the amino acid residues of positions 1, 6, 10, 13, 14, 17, 18, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 are given). All values are given in %. For example, at position 13, the human IFN-β amino acid is serine (see FIG. 1) (SEQ ID NO: 1); in IbA proteins, 67.7% of the top 1000 sequences had phenylalanine at this position and 31.4% of the sequences had tyrosine. None of the IbA sequences had serine at this position. Similarly, at position 118, the human IFN-β amino acid is serine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 89.1 % of the top 1000 sequences had alanine at this position and 10.9% of the sequences had tyrosine. None of the IbA sequences had serine at this position.

FIG. 6B (SEQ ID NO:6) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 3 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 6C and FIG. 6D (SEQ ID NO:7–8) depict preferred IbA sequences based on the PDA analysis of IFN-β A-chain core 3 sequence, generated not only by the direct MC calculation following DEE, but also those after cleaning the MC list (C) and when running MC over the complete sequence space starting from the ground state generated by the direct MC calculation (D). Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 7A depicts the mutation pattern of IFN-β A-chain core 4 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 4 sequences. See FIG. 6A for details of figure legend. For example, at position 17, the human IFN-β amino acid is cysteine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 82.9% of the top 1000 sequences had aspartic acid at this position, 7.1% had threonine, 4.5% had alanine, 4.1% had leucine and 1.4% had valine. None of the IbA sequences had cysteine at this position.

FIG. 7B (SEQ ID NO:9) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 4 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 7C and FIG. 7D (SEQ ID NOS:10–11) depict preferred IbA sequences based on the PDA analysis of IFN-β

A-chain core 4 sequence, generated not only by the direct MC calculation following DEE, but also those after cleaning the MC list (C) and when running MC over the complete sequence space starting from the ground state generated by the direct MC calculation (D). Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 8A depicts the mutation pattern of IFN-β A-chain core 5 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 5 sequences. See FIG. 6A for details of figure legend. For example, at position 84, the human IFN-β amino acid is valine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 99.5% of the top 1000 sequences had isoleucine at this position and 0.5% had leucine. None of the IbA sequences had valine at this position.

FIG. 8B (SEQ ID NO:12) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 5 sequence. Amino acid residues different from the human IFN-p (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 8C and FIG. 8D (SEQ ID NOS:13–14) depict preferred IbA sequences based on the PDA analysis of IFN-β A-chain core 5 sequence, generated not only by the direct MC calculation following DEE, but also those after cleaning the MC list (C) and when running MC over the complete sequence space starting from the ground state generated by the direct MC calculation (D). Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 9A depicts the mutation pattern of IFN-β A-chain core 6 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of A-chain IFN-β core 6 sequences. See FIG. 6A for details of figure legend. For example, at position 118, the human IFN-β amino acid is serine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 100% of the top 1000 sequences had alanine. None of the IbA sequences had serine at this position.

FIG. 9B (SEQ ID NO:15) depicts a preferred IbA sequence based on the PDA analysis of IFN-β A-chain core 6 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 9C and FIG. 9D (SEQ ID NOS:16–17) depict preferred IbA sequences based on the PDA analysis of IFN-β A-chain core 6 sequence, generated not only by the direct MC calculation following DEE, but also those after cleaning the MC list (C) and when running MC over the complete sequence space starting from the ground state generated by the direct MC calculation (D). Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 10A depicts the mutation pattern of IFN-β B-chain core 1 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 1 sequences (only the amino acid residues of positions 6, 21, 55, 56, 59, 62, 63, 66, 69, 84, 87, 91, 98, 122, 129, 133, 146, 150, 157, and 160 are given). All values are given in %. For example, at position 87, the human IFN-β amino acid is leucine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 74.6% of the top 1000 sequences had phenylalanine at this position, and only 21.5% of the sequences had leucine. Similarly, for position 84 (valine in human IFN-β), isoleucine (62.3%) is preferred over valine (25.4%).

FIG. 10B (SEQ ID NO:18) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 1 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) SEQ ID NO:1) are shown in bold and are underlined.

FIG. 11A depicts the mutation pattern of IFN-β B-chain core 2 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 2 sequences (only the amino acid residues of positions 1, 6, 10, 14, 17, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 81, 84, 87, 91, 94, 95, 98, 102, 115, 122, 125, 126, 129, 130, 133, 138, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 are given). All values are given in %. For example, at position 56, the human IFN-β amino acid is alanine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 97.6% of the top 1000 sequences had leucine at this position, and only 2.4% of the sequences had alanine. Similarly, for position 91 (valine in human IFN-β), isoleucine (68.5%) and leucine (27.7%) are preferred over valine (3.8%).

FIG. 11B (SEQ ID NO:19) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 2 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 12A depicts the mutation pattern of IFN-β B-chain core 3 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 3 sequences (only the amino acid residues of positions 1, 6, 10, 13, 14, 15, 17, 21, 38, 50, 55, 56, 58, 10 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 are given). All values are given in %. For example, at position 13, the human IFN-β amino acid is serine (see FIG. 1); in IbA proteins, 92.7% of the top 1000 sequences had leucine at this position and 7.3% of the sequences had alanine. None of the IbA sequences had serine at this position. Similarly, at position 118, the human IFN-β amino acid is serine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 100% of the top 1000 sequences had leucine at this position.

FIG. 12B (SEQ ID NO:20) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 3 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 13A depicts the mutation pattern of IFN-β B-chain core 4 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 4 sequences. See FIG. 12A for details of figure legend. For example, at position 56, the human IFN-β amino acid is alanine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 97.7% of the top 1000 sequences had leucine at this position and only 2.3% had alanine. Similarly, at position 114, the human IFN-β amino acid is glycine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 100% of the top 1000 sequences had phenylalanine at this position.

FIG. 13B (SEQ ID NO:21) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 4 sequence. Amino acid residues different from the human IFN-β (see FIG. 1) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. 14A depicts the mutation pattern of IFN-β B-chain core 5 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 5 sequences (only the amino acid residues of positions 1, 6, 10, 13, 14, 17, 18, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 are given). For example, at position 56, the human IFN-β amino acid is alanine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 97.6% of the top 1000 sequences had leucine at this position and only 2.4% had alanine. Similarly, at position 114, the human IFN-β amino acid is glycine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 100% of the top 1000 sequences had leucine at this position.

FIG. **14B** (SEQ ID NO:1) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 5 sequence. Amino acid residues different from the human IFN-β (see FIG. **1**) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. **15A** depicts the mutation pattern of IFN-β B-chain core 6 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 6 sequences. See FIG. 14A for details of figure legend. For example, at position 118, the human IFN-β amino acid is serine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 99.4% of the top 1000 sequences had glutamic acid at this position and 0.6% had alanine. None of the IbA sequences had serine at this position. Similarly, for position 161 (threonine in human IFN-β), glutamic acid (86.4%) is preferred over threonine (12.1 %).

FIG. **15B** (SEQ ID NO:23) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 6 sequence. Amino acid residues different from the human IFN-β (see FIG. **1**) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. **16A** depicts the mutation pattern of IFN-β, B-chain core 7 sequences based on the analysis of the lowest 1000 protein sequences generated by Monte Carlo analysis of B-chain IFN-β core 7 sequences. See FIG. 14A for details of figure legend. For example, at position 17, the human IFN-β amino acid is cysteine (see FIG. 1) (SEQ ID NO:1); in IbA proteins, 32.8% of the top 1000 sequences had threonine at this position, 31 % had alanine, 29% had aspartic acid, 5% had glutamic acid, 1.4% had serine, and 0.8% had glycine. None of the IbA sequences had cysteine at this position.

FIG. **16B** (SEQ ID NO:24) depicts a preferred IbA sequence based on the PDA analysis of IFN-β B-chain core 7 sequence. Amino acid residues different from the human IFN-β (see FIG. **1**) (SEQ ID NO:1) are shown in bold and are underlined.

FIG. **17** depicts the synthesis of a full-length gene and all possible mutations by PCR. Overlapping oligonucleotides corresponding to the full-length gene (black bar, Step 1) and comprising one or more desired mutations are synthesized, heated and annealed. Addition of DNA polymerase to the annealed oligonucleotdes results in the 5' to 3' synthesis of DNA (Step 2) to produce longer DNA fragments (Step 3). Repeated cycles of heating, annealing, and DNA synthesis (Step 4) result in the production of longer DNA, including some full-length molecules. These can be selected by a second round of PCR using primers (indicated by arrows) corresponding to the end of the full-length gene (Step 5).

FIG. **18** depicts a preferred scheme for synthesizing an IbA library of the invention. The wild type gene, or any starting gene, such as the gene for the global minima gene, can be used. Oligonucleotides comprising sequences that encode different amino acids at the different variant positions (indicated in the Figure by box 1, box 2, and box 3) can be used during PCR. Those primers can be used in combination with standard primers. This generally requires fewer oligonucleotides and can result in fewer errors.

FIG. **19** depicts an overlapping extension method. At the top of FIG. **19** is the template DNA showing the locations of the regions to be mutated (black boxes) and the binding sites of the relevant primers (arrows). The primers R1 and R2 represent a pool of primers, each containing a different mutation; as described herein, this may be done using different ratios of primers if desired. The variant position is flanked by regions of homology sufficient to get hybridization. Thus, as shown in this example, oligos R1 and F2 comprise a region of homology and so do oligos R2 and F3. In this example, three separate PCR reactions are done for step 1. The first reaction contains the template plus oligos F1 and R1. The second reaction contains template plus oligos F2 and R2, and the third contains the template and oligos F3 and R3. The reaction products are shown. In Step 2, the products from Step 1 tube 1 and Step 1 tube 2 are taken. After purification away from the primers, these are added to a fresh PCR reaction together with F1 and R4. During the denaturation phase of the PCR, the overlapping regions anneal and the second strand is synthesized. The product is then amplified by the outside primers, F1 and R4. In Step 3, the purified product from Step 2 is used in a third PCR reaction, together with the product of Step 1, tube 3 and the primers F1 and R3. The final product corresponds to the full length gene and contains the required mutations. Alternatively, Step 2 and Step 3 can be performed in one PCR reaction.

FIG. **20** depicts a ligation of PCR reaction products to synthesize the libraries of the invention. In this technique, the primers also contain an endonuclease restriction site (RE), either generating blunt ends, 5' overhanging ends or 3' overhanging ends. We set up three separate PCR reactions for Step 1. The first reaction contains the template plus oligos F1 and R1. The second reaction contains template plus oligos F2 and R2, and the third contains the template and oligos F3 and R3. The reaction products are shown. In Step 2, the products of Step 1 are purified and then digested with the appropriate restriction endonuclease. The digestion products from Step 2, tube 1 and Step 2, tube 2 are ligated together with DNA ligase (Step 3). The products are then amplified in Step 4 using oligos F1 and R4. The whole process is then repeated by digesting the amplified products, ligating them to the digested products of Step 2, tube 3, and then amplifying the final product using oligos F1 and R3. It would also be possible to ligate all three PCR products from Step 1 together in one reaction, providing the two restriction sites (RE1 and RE2) were different.

FIG. **21** depicts blunt end ligation of PCR products. In this technique, oligos such as F2 and R1 or R2 and F3 do not overlap, but they abut. Again three separate PCR reactions are performed. The products from·tube 1 and tube 2 (see FIG. **20**, Step 1) are ligated, and then amplified with outside primers F1 and R4. This product is then ligated with the product from Step 1, tube 3. The final products are then amplified with primers F1 and R3.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to novel proteins and nucleic acids possessing interferon-beta activity (sometimes referred to herein as "IbA proteins" and "IbA nucleic acids"). The proteins are generated using a system previously described in WO98/47089 and U.S. Ser. Nos. 09/058,459, 09/127,926, 60/104,612, 60/158,700, 09/419,351, 60/181,630, 60/186,904, and U.S patent application, entitled *Protein Design Automation For Protein Libraries* (Filed: Apr. 14, 2000; Inventor: Bassil Dahiyat), all of which are expressly incorporated by reference in their entirety, that is a computational modeling system that allows the generation of extremely stable proteins without necessarily disturbing the biological functions of the protein itself. In this way, novel IbA proteins and nucleic acids are generated, that can have a plurality of mutations in comparison to the wild-type enzyme yet retain significant activity.

Generally, there are a variety of computational methods that can be used to generate the IbA proteins of the invention. In a preferred embodiment, sequence based methods are used. Alternatively, structure based methods, such as PDA, described in detail below, are used.

Similarly, molecular dynamics calculations can be used to computationally screen sequences by individually calculating mutant sequence scores and compiling a rank ordered list.

In a preferred embodiment, residue pair potentials can be used to score sequences (Miyazawa et al., Macromolecules 18(3):534–552 (1985), expressly incorporated by reference) during computational screening.

In a preferred embodiment, sequence profile scores (Bowie et al., Science 253(5016):164–70 (1991), incorporated by reference) and/or potentials of mean force (Hendlich et al., J. Mol. Biol. 216(1):167–180 (1990), also incorporated by reference) can also be calculated to score sequences. These methods assess the match between a sequence and a 3D protein structure and hence can act to screen for fidelity to the protein structure. By using different scoring functions to rank sequences, different regions of sequence space can be sampled in the computational screen.

Furthermore, scoring functions can be used to screen for sequences that would create metal or co-factor binding sites in the protein (Hellinga, Fold Des. 3(1):R1-8 (1998), hereby expressly incorporated by reference). Similarly, scoring functions can be used to screen for sequences that would create disulfide bonds in the protein. These potentials attempt to specifically modify a protein structure to introduce a new structural motif.

In a preferred embodiment, sequence and/or structural alignment programs can be used to generate the IbA proteins of the invention. As is known in the art, there are a number of sequence-based alignment programs; including for example, Smith-Waterman searches, Needleman-Wunsch, Double Affine Smith-Waterman, frame search, Gribskov/GCG profile search, Gribskov/GCG profile scan, profile frame search, Bucher generalized profiles, Hidden Markov models, Hframe, Double Frame, Blast, Psi-Blast, Clustal, and GeneWise.

As is known in the art, there are a number of sequence alignment methodologies that can be used. For example, sequence homology based alignment methods can be used to create sequence alignments of proteins related to the target structure (Altschul et al., J. Mol. Biol. 215(3):403–410 (1990), Altschul et al., Nucleic Acids Res. 25:3389–3402 (1997), both incorporated by reference). These sequence alignments are then examined to determine the observed sequence variations. These sequence variations are tabulated to define a set of IbA proteins.

Sequence based alignments can be used in a variety of ways. For example, a number of related proteins can be aligned, as is known in the art, and the "variable" and "conserved" residues defined; that is, the residues that vary or remain identical between the family members can be defined. These results can be used to generate a probability table, as outlined below. Similarly, these sequence variations can be tabulated and a secondary library defined from them as defined below. Alternatively, the allowed sequence variations can be used to define the amino acids considered at each position during the computational screening. Another variation is to bias the score for amino acids that occur in the sequence alignment, thereby increasing the likelihood that they are found during computational screening but still allowing consideration of other amino acids. This bias would result in a focused library of IbA proteins but would not eliminate from consideration amino acids not found in the alignment. In addition, a number of other types of bias may be introduced. For example, diversity may be forced; that is, a "conserved" residue is chosen and altered to force diversity on the protein and thus sample a greater portion of the sequence space. Alternatively, the positions of high variability between family members (i.e. low conservation) can be randomized, either using all or a subset of amino acids. Similarly, outlier residues, either positional outliers or side chain outliers, may be eliminated.

Similarly, structural alignment of structurally related proteins can be done to generate sequence alignments (Orengo et al., Structure 5(8):1093–108 (1997); Holm et al., Nucleic Acids Res. 26(1):316–9 (1998), both of which are incorporated by reference). These sequence alignments can then be examined to determine the observed sequence variations. Libraries can be generated by predicting secondary structure from sequence, and then selecting sequences that are compatible with the predicted secondary structure. There are a number of secondary structure prediction methods such as helix-coil transition theory (Munoz and Serrano, Biopolymers 41:495, 1997), neural networks, local structure alignment and others (e.g., see in Selbig et al., Bioinformatics 15:1039–46, 1999).

Similarly, as outlined above, other computational methods are known, including, but not limited to, sequence profiling [Bowie and Eisenberg, Science 253(5016):164–70, (1991)], rotamer library selections [Dahiyat and Mayo, Protein Sci. 5(5):895–903 (1996); Dahiyat and Mayo, Science 278(5335):82–7 (1997); Desjarlais and Handel, Protein Science 4:2006–2018 (1995); Harbury et Proc. Natl. Acad. Sci. U.S.A. 92(18):8408–8412 (1995); Kono et al., Proteins: Structure, Function and Genetics 19:244–255 (1994); Hellinga and Richards, Proc. Natl. Acad. Sci. U.S.A. 91:5803–5807 (1994)]; and residue pair potentials [Jones, Protein Science 3: 567–574, (1994)]; PROSA [Heindlich et al., J. Mol. Biol. 216:167–180 (1990)]; THREADER [Jones et al., Nature 358:86–89 (1992)], and other inverse folding methods such as those described by Simons et al. [Proteins, 34:535–543, (1999)], Levitt and Gerstein [Proc. Natl. Acad. Sci. U.S.A., 95:5913–5920, (1998)], Godzik and Skolnick [Proc. Natl. Acad. Sci. U.S.A., 89:12098–102, (1992)], Godzik et al. [J. Mol. Biol. 227:227–38, (1992)], and other profile methods [Gribskov et al. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358 (1987) and Fischer and Eisenberg, Protein Sci. 5:947–955 (1996), Rice and Eisenberg J. Mol. Biol. 267:1026–1038(1997)], all of which are expressly incorporated by reference. In addition, other computational methods such as those described by Koehl and Levitt (J. Mol. Biol. 293:1161–1181 (1999); J. Mol. Biol. 293:1183–1193 (1999); expressly incorporated by reference) can be used to create a protein sequence library which can optionally then be used to generate a smaller secondary library for use in experimental screening for improved properties and function. In addition, there are computational methods based on forcefield calculations such as SCMF that can be used as well for SCMF, see Delarue et al. Pac. Symp. Biocomput. 109–21 (1997); Koehl et al., J. Mol. Biol. 239:249–75 (1994); Koehl et al., Nat. Struct. Biol. 2:163–70 (1995); Koehl et al., Curr. Opin. Struct. Biol. 6:222–6 (1996); Koehl et al., J. Biol. 293:1183–93 (1999); Koehl et al., J. Mol. Biol. 293:1161–81 (1999); Lee J., Mol. Biol.236:918–39 (1994); and Vasquez Biopolymers 36:53–70 (1995); all of which are expressly incorporated by reference. Other forcefield calculations that can be used to optimize the conformation of a sequence within a computational method, or to generate de novo optimized sequences as outlined herein include, but are not limited to, OPLS-AA [Jorgensen et al., J. Am. Chem. Soc. 118:11225–11236 (1996); Jorgensen, W. L.; BOSS, Version 4.1; Yale University: New Haven, CT (1999)]; OPLS [Jorgensen et al., J. Am. Chem. Soc.110:1657ff (1988); Jorgensen et al., J. Am. Chem. Soc.112:4768ff (1990)]; UNRES (United Residue Forcefield; Liwo et al., Protein Science 2:1697–1714 (1993); Liwo et al., Protein Science 2:1715–1731 (1993); Liwo et al., J. Comp. Chem. 18:849–873 (1997); Liwo et al. Comp. Chem. 18:874–884 (1997); Liwo et al., J. Comp. Chem. 19:259–276 (1998); Forcefield for Protein Structure Prediction (Liwo et al., Proc. Natl. Acad. Sci. U.S.A. 96:5482–5485 (1999)]; ECEPP/3 [Liwo et al., J Protein Chem. 13(4):375–80 (1994)]; AMBER 1.1 force field (Weiner et a Am. Chem. Soc. 106:765–784); AMBER 3.0 force field [U. C. Singh et al., Proc. Natl. Acad. Sci. U.S.A. 82:755–759 (1985)]; CHARMM and CHARMM22 (Brooks et al., J. Comp. Chem. 4:187–217); cvff3.0 [Dauber-Osguthorpe et al., Proteins: Structure, Function and Genetics, 4:31–47 (1988)]; cff99:1 (Maple et al., J. Comp. Chem. 15:162–182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego Calif.) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego Calif.), all of which are expressly incorporated by reference. In fact, as is outlined below, these forcefield methods may be used to generate the secondary library directly; that is, no primary library is generated; rather, these methods can be used to generate a probability table from which the secondary library is directly generated.

In a preferred embodiment, the computational method used to generate the primary library is Protein Design Automation (PDA), as is described in U.S. Ser. Nos. 60/061, 097, 60/043,464, 60/054,678, 09/127,926, 60/104,612, 60/158,700, 09/419,351, 60/181,630, 60/186,904, U.S patent application, entitled *Protein Design Automation For Protein Libraries* (Filed: Apr. 14, 2000; Inventor: Bassil Dahiyat) and PCT US98/07254, all of which are expressly incorporated herein by reference. Briefly, PDA can be described as follows. A known protein structure is used as the starting point. The residues to be optimized are then identified, which may be the entire sequence or subset(s) thereof. The side chains of any positions to be varied are then removed. The resulting structure consisting of the protein backbone and the remaining sidechains is called the template. Each variable residue position is then preferably classified as a core residue, a surface residue, or a boundary residue; each classification defines a subset of possible amino acid residues for the position (for example, core residues generally will be selected from the set of hydrophobic residues, surface residues generally will be selected from the hydrophilic residues, and boundary residues may be either). Each amino acid can be represented by a discrete set of all allowed conformers of each side chain, called rotamers. Thus, to arrive at an optimal sequence for a backbone, all possible sequences of rotamers must be screened, where each backbone position can be occupied either by each amino acid in all its possible rotameric states, or a subset of amino acids, and thus a subset of rotamers.

Two sets of interactions are then calculated for each rotamer at every position: the interaction of the rotamer side chain with all or part of the backbone (the "singles" energy, also called the rotamer/template or rotamer/backbone energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position or a subset of the other positions (the "doubles" energy, also called the rotamer/rotamer energy). The energy of each of these interactions is calculated through the use of a variety of scoring functions, which include the energy of van der Waal's forces, the energy of hydrogen bonding, the energy of secondary structure propensity, the energy of surface area solvation and the electrostatics. Thus, the total energy of each rotamer interaction, both with the backbone and other rotamers, is calculated, and stored in a matrix form.

The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences, a number which grows exponentially with sequence length and renders the calculations either unwieldy or impossible in real time. Accordingly, to solve this combinatorial search problem, a "Dead End Elimination" (DEE) calculation is performed. The DEE calculation is based on the fact that if the worst total interaction of a first rotamer is still better than the best total interaction of a second rotamer, then the second rotamer cannot be part of the global optimum solution. Since the energies of all rotamers have already been calculated, the DEE approach only requires sums over the sequence length to test and eliminate rotamers, which speeds up the calculations considerably. DEE can be rerun comparing pairs of rotamers, or combinations of rotamers, which will eventually result in the determination of a single sequence which represents the global optimum energy.

Once the global solution has been found, a Monte Carlo search may be done to generate a rank-ordered list of sequences in the neighborhood of the DEE solution. Starting at the DEE solution, random positions are changed to other rotamers, and the new sequence energy is calculated. If the new sequence meets the criteria for acceptance, it is used as a starting point for another jump. After a predetermined number of jumps, a rank-ordered list of sequences is generated. Monte Carlo searching is a sampling technique to explore sequence space around the global minimum or to find new local minima distant in sequence space. As is more additionally outlined below, there are other sampling techniques that can be used, including Boltzman sampling, genetic algorithm techniques and simulated annealing. In addition, for all the sampling techniques, the kinds of jumps allowed can be altered (e.g. random jumps to random residues, biased jumps (to or away from wild-type, for example), jumps to biased residues (to or away from similar residues, for example), etc.). Similarly, for all the sampling techniques, the acceptance criteria of whether a sampling jump is accepted can be altered.

As outlined in U.S. Ser. No. 09/127,926, the protein backbone (comprising (for a naturally occuring protein) the nitrogen, the carbonyl carbon, the α-carbon, and the carbonyl oxygen, along with the direction of the vector from the α-carbon to the β-carbon) may be altered prior to the computational analysis, by varying a set of parameters called supersecondary structure parameters.

Once a protein structure backbone is generated (with alterations, as outlined above) and input into the computer, explicit hydrogens are added if not included within the structure (for example, if the structure was generated by X-ray crystallography, hydrogens must be added). After hydrogen addition, energy minimization of the structure is run, to relax the hydrogens as well as the other atoms, bond angles and bond lengths. In a preferred embodiment, this is done by doing a number of steps of conjugate gradient minimization [Mayo et al., J. Phys. Chem. 94:8897 (1990)] of atomic coordinate positions to minimize the Dreiding force field with no electrostatics. Generally from about 10 to about 250 steps is preferred, with about 50 being most preferred.

The protein backbone structure contains at least one variable residue position. As is known in the art, the residues, or amino acids, of proteins are generally sequentially numbered starting with the N-terminus of the protein. Thus a protein having a methionine at it's N-terminus is said to have a methionine at residue or amino acid position 1, with the next residues as 2, 3, 4, etc. At each position, the wild type (i.e. naturally occuring) protein may have one of at least 20 amino acids, in any number of rotamers. By "variable residue position" herein is meant an amino acid position of the protein to be designed that is not fixed in the design method as a specific residue or rotamer, generally the wild-type residue or rotamer.

In a preferred embodiment, all of the residue positions of the protein are variable. That is, every amino acid side chain may be altered in the methods of the present invention. This is particularly desirable for smaller proteins, although the present methods allow the design of larger proteins as well. While there is no theoretical limit to the length of the protein which may be designed this way, there is a practical computational limit.

In an alternate preferred embodiment, only some of the residue positions of the protein are variable, and the remainder are "fixed", that is, they are identified in the three dimensional structure as being in a set conformation. In some embodiments, a fixed position is left in its original conformation (which may or may not correlate to a specific rotamer of the rotamer library being used). Alternatively, residues may be fixed as a non-wild type residue; for example, when known site-directed mutagenesis techniques have shown that a particular residue is desirable (for example, to eliminate a proteolytic site or alter the substrate specificity of an enzyme), the residue may be fixed as a particular amino acid. Alternatively, the methods of the present invention may be used to evaluate mutations de novo, as is discussed below. In an alternate preferred embodiment, a fixed position may be "floated"; the amino acid at that position is fixed, but different rotamers of that amino acid are tested. In this embodiment, the variable residues may be at least one, or anywhere from 0.1% to 99.9% of the total number of residues. Thus, for example, it may be possible to change only a few (or one) residues, or most of the residues, with all possibilities in between.

In a preferred embodiment, residues which can be fixed include, but are not limited to, structurally or biologically functional residues; alternatively, biologically functional residues may specifically not be fixed. For example, residues which are known to be important for biological activity, such as the residues which the binding site for a binding partner (ligand/receptor, antigen/antibody, etc.), phosphorylation or glycosylation sites which are crucial to biological function, or structurally important residues, such as disulfide bridges, metal binding sites, critical hydrogen bonding residues, residues critical for backbone conformation such as proline or glycine, residues critical for packing interactions, etc. may all be fixed in their amino acid identity and a single rotamer conformation, or "floated", which only fixes the identity but not the rotamer conformation.

Similarly, residues which may be chosen as variable residues may be those that confer undesirable biological attributes, such as susceptibility to proteolytic degradation, dimerization or aggregation sites, glycosylabon sites which may lead to immune responses, unwanted binding activity, unwanted allostery, undesirable enzyme activity but with a preservation of binding, etc.

In a preferred embodiment, each variable position is classified as either a core, surface or boundary residue position, although in some cases, as explained below, the variable position may be set to glycine to minimize backbone strain. In addition, as outlined herein, residues need not be classified, they can be chosen as variable and any set of amino acids may be used. Any combination of core, surface and boundary positions can be utilized: core, surface and boundary residues; core and surface residues; core and boundary residues, and surface and boundary residues, as well as core residues alone, surface residues alone, or boundary residues alone.

The classification of residue positions as core, surface or boundary may be done in several ways, as will be appreciated by those in the art. In a preferred embodiment, the classification is done via a visual scan of the original protein backbone structure, including the side chains, and assigning a classification based on a subjective evaluation of one skilled in the art of protein modelling. Alternatively, a preferred embodiment utilizes an assessment of the orientation of the Cα-Cβ vectors relative to a solvent accessible surface computed using only the template Cα atoms, as outlined in U.S. Ser. Nos. 60/061,097, 60/043,464, 60/054, 678, 09/127,926 60/104,612, 60/158,700, 09/419,351, 60/181,630, 60/186,904, U.S patent application, entitled *Protein Design Automation For Protein Libraries* (Filed: Apr. 14, 2000; Inventor: Bassil Dahiyat) and PCT US98/07254. Alternatively, a surface area calculation can be done.

Suitable core and boundary positions for IbA proteins are outlined below.

Once each variable position is classified as either core, surface or boundary, a set of amino acid side chains, and thus a set of rotamers, is assigned to each position. That is, the set of possible amino acid side chains that the program will allow to be considered at any particular position is chosen. Subsequently, once the possible amino acid side chains are chosen, the set of rotamers that will be evaluated at a particular position can be determined. Thus, a core residue will generally be selected from the group of hydrophobic residues consisting of alanine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine (in some embodiments, when the a scaling factor of the van der Waals scoring function, described below, is low, methionine is removed from the set), and the rotamer set for each core position potentially includes rotamers for these eight amino acid side chains (all the rotamers if a backbone independent library is used, and subsets if a rotamer dependent backbone is used). Similarly, surface positions are generally selected from the group of hydrophilic residues consisting of alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine and histidine. The rotamer set for each surface position thus includes rotamers for these ten residues. Finally, boundary positions are generally chosen from alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine histidine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine. The rotamer set for each boundary position thus potentially includes every rotamer for these seventeen residues (assuming cysteine, glycine and proline are not used, although they can be). Additionally, in some preferred embodiments, a set of 18 naturally occuring amino acids (all except cysteine and proline, which are known to be particularly disruptive) are used.

Thus, as will be appreciated by those in the art, there is a computational benefit to classifying the residue positions, as it decreases the number of calculations. It should also be noted that there may be situations where the sets of core, boundary and surface residues are altered from those described above; for example, under some circumstances, one or more amino acids is either added or subtracted from the set of allowed amino acids. For example, some proteins which dimerize or multmerize, or have ligand binding sites, may contain hydrophobic surface residues, etc. In addition, residues that do not allow helix "capping" or the favorable interaction with an a-helix dipole may be subtracted from a set of allowed residues. This modification of amino acid groups is done on a residue by residue basis.

In a preferred embodiment, proline, cysteine and glycine are not included in the list of possible amino acid side chains, and thus the rotamers for these side chains are not used. However, in a preferred embodiment, when the variable residue position has a ϕ angle (that is, the dihedral angle defined by 1) the carbonyl carbon of the preceding amino acid; 2) the nitrogen atom of the current residue; 3) the α-carbon of the current residue; and 4) the carbonyl carbon of the current residue) greater than 0°, the position is set to glycine to minimize backbone strain.

Once the group of potential rotamers is assigned for each variable residue position, processing proceeds as outlined in U.S. Ser. No. 09/127, 926 and PCT US98/07254. This processing step entails analyzing interactions of the rotamers with each other and with the protein backbone to generate optimized protein sequences. Simplistically, the processing initially comprises the use of a number of scoring functions to calculate energies of interactions of the rotamers, either to the backbone itself or other rotamers. Preferred PDA scoring functions include, but are not limited to, a Van der Waals potential scoring function, a hydrogen bond potential scoring function, an atomic solvation scoring function, a secondary structure propensity scoring function and an electrostatic scoring function. As is further described below, at least one scoring function is used to score each position, although the scoring functions may differ depending on the position classification or other considerations, like favorable interaction with an α-helix dipole. As outlined below, the total energy which is used in the calculations is the sum of the energy of each scoring function used at a particular position, as is generally shown in Equation 1:

$$E_{total}=nE_{vdw}+nE_{as}+nE_{h\text{-}bonding}+nE_{ss}+nE_{elec} \qquad \text{Equation 1}$$

In Equation 1, the total energy is the sum of the energy of the van der Waals potential (E_{vdw}), the energy of atomic solvation (E_{as}), the energy of hydrogen bonding ($E_{h\text{-}bonding}$), the energy of secondary structure (E_{ss}) and the energy of electrostatic interaction (E_{elec}). The term n is either 0 or 1, depending on whether the term is to be considered for the particular residue position.

As outlined in U.S. Ser. Nos. 60/061,097, 60/043,464, 60/054,678, 09/127,926, 60/104,612, 60/158,700, 09/419, 351, 60/181,630, 60/186,904, U.S patent application, entitled *Protein Design Automation For Protein Libraries* (Filed: Apr. 14, 2000; Inventor: Bassil Dahiyat) and PCT US98/07254, any combination of these scoring functions, either alone or in combination, may be used. Once the scoring functions to be used are identified for each variable position, the preferred first step in the computational analysis comprises the determination of the interaction of each possible rotamer with all or part of the remainder of the protein. That is, the energy of interaction, as measured by one or more of the scoring functions, of each possible rotamer at each variable residue position with either the backbone or other rotamers, is calculated. In a preferred embodiment, the interaction of each rotamer with the entire remainder of the protein, i.e. both the entire template and all other rotamers, is done. However, as outlined above, it is possible to only model a portion of a protein, for example a domain of a larger protein, and thus in some cases, not all of the protein need be considered. The term "portion", or similar grammatical equivalents thereof, as used herein, with regard to a protein refers to a fragment of that protein. This fragment may range in size from 5–10 amino acid residues to the entire amino acid sequence minus one amino acid. Accordingly, the term "portion", as used herein, with regard to a nucleic refers to a fragment of that nucleic acid. This fragment may range in size from 6–10 nucleotides to the entire nucleic acid sequence minus one nucleotide.

In a preferred embodiment, the first step of the computational processing is done by calculating two sets of interactions for each rotamer at every position: the interaction of the rotamer side chain with the template or backbone (the "singles" energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position (the "doubles" energy), whether that position is varied or floated. It should be understood that the backbone in this case includes both the atoms of the protein structure backbone, as well as the atoms of any fixed residues, wherein the fixed residues are defined as a particular conformation of an amino acid.

Thus, "singles" (rotamer/template) energies are calculated for the interaction of every possible rotamer at every variable residue position with the backbone, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the rotamer and every hydrogen bonding atom of the backbone is evaluated, and the E_{HB} is calculated for each possible rotamer at every variable position. Similarly, for the van der Waals scoring function, every atom of the rotamer is compared to every atom of the template (generally excluding the backbone atoms of its own residue), and the E_{vdW} is calculated for each possible rotamer at every variable residue position. In addition, generally no van der Waals energy is calculated if the atoms are connected by three bonds or less. For the atomic solvation scoring function, the surface of the rotamer is measured against the surface of the template, and the E_{as} for each possible rotamer at every variable residue position is calculated. The secondary structure propensity scoring function is also considered as a singles energy, and thus the total singles energy may contain an E_{ss} term. As will be appreciated by those in the art, many of these energy terms will be close to zero, depending on the physical distance between the rotamer and the template position; that is, the farther apart the two moieties, the lower the energy.

For the calculation of "doubles" energy (rotamer/rotamer), the interaction energy of each possible rotamer is compared with every possible rotamer at all other variable residue positions. Thus, "doubles" energies are calculated for the interaction of every possible rotamer at every variable residue position with every possible rotamer at every other variable residue position, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the first rotamer and every hydrogen bonding atom of every possible second rotamer is evaluated, and the E_{HB} is calculated for each possible rotamer pair for any two variable positions. Similarly, for the van der Waals scoring function, every atom of the first rotamer is compared to every atom of every possible second rotamer, and the E_{vdw} is calculated for each possible rotamer pair at every two variable residue positions. For the atomic solvation scoring function, the surface of the first rotamer is measured against the surface of every possible second rotamer, and the E_{as} for each possible rotamer pair at every two variable residue positions is calculated. The secondary structure propensity scoring function need not be run as a "doubles" energy, as it is considered as a component of the "singles" energy. As will be appreciated by those in the art, many of these double energy terms will be close to zero, depending on the physical distance between the first rotamer and the second rotamer; that is, the farther apart the two moieties, the lower the energy.

In addition, as will be appreciated by those in the art, a variety of force fields that can be used in the PDA calculations can be used, including, but not limited to, Dreiding I and Dreiding II [Mayo et al, J. Phys. Chem. 94:8897 (1990)], AMBER [Weiner et al., J. Amer. Chem. Soc. 106:765 (1984) and Weiner et al., J. Comp. Chem. 106:230 (1986)], MM2 [Allinger, J. Chem. Soc. 99:8127 (1977), Liljefors et al., J. Com. Chem. 8:1051 (1987)]; MMP2 [Sprague et al., J. Comp. Chem. 8:581 (1987)]; CHARMM [Brooks et al., J. Comp. Chem. 106:187 (1983)]; GROMOS; and MM3 [Allinger et al., J. Amer. Chem. Soc. 111:8551 (1989)], OPLS-AA [Jorgensen et al., J. Am. Chem. Soc. 118:11225–11236 (1996); Jorgensen, W. L.; BOSS, Version 4.1; Yale University: New Haven, Conn. (1999)]; OPLS [Jorgensen et al., J. Am. Chem. Soc.110:1657ff (1988); Jorgensen et al., J Am. Chem. Soc. 112:4768ff (1990)]; UNRES (United Residue Forcefield; Liwo et al., Protein Science 2:1697–1714 (1993); Liwo et al., Protein Science 2:1715–1731 (1993); Liwo et al., J. Comp. Chem. 18:849–873 (1997); Liwo et al., J. Comp. Chem. 18:874–884 (1997); Liwo et al., J. Comp. Chem. 19:259–276 (1998); Forcefield for Protein Structure Prediction (Liwo et al., Proc. Natl. Acad. Sci. U.S.A 96:5482–5485 (1999)]; ECEPP/3 [Liwo et al., J Protein Chem. 13(4) :375–80 (1994)]; A field (Weiner, et al., J. Am. Chem. Soc. 106:765–784); AMBER 3.0 force field (U.C. Singh et al., Proc. Natl. Acad. Sci. U.S.A. 82:755–759); CHARMM and CHARMM22 (Brooks et al., J. Comp. Chem. 4:187–217); cvff3.0 [Dauber-Osguthorpe, et al., Proteins: Structure, Function and Genetics, 4:31–47 (1988)]; cff91 (Maple, et al., J. Comp. Chem. 15:162–182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego Calif.) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego Calif.), all of which are expressly incorporated by reference.

Once the singles and doubles energies are calculated and stored, the next step of the computational processing may occur. As outlined in U.S. Ser. No. 09/127,926 and PCT US98/07254, preferred embodiments utilize a Dead End Elimination (DEE) step, and preferably a Monte Carlo step.

PDA, viewed broadly, has three components that may be varied to alter the output (e.g. the primary library): the scoring functions used in the process; the filtering technique, and the sampling technique.

In a preferred embodiment, the scoring functions may be altered. In a preferred embodiment, the scoring functions outlined above may be biased or weighted in a variety of ways. For example, a bias towards or away from a reference sequence or family of sequences can be done; for example, a bias towards wild-type or homolog residues may be used. Similarly, the entire protein or a fragment of it may be biased; for example, the active site may be biased towards wild-type residues, or domain residues towards a particular desired physical property can be done. Furthermore, a bias towards or against increased energy can be generated. Additional scoring function biases include, but are not limited to applying electrostatic potential gradients or hydrophobicity gradients, adding a substrate or binding partner to the calculation, or biasing towards a desired charge or hydrophobicity.

In addition, in an alternative embodiment, there are a variety of additional scoring functions that may be used. Additional scoring functions include, but are not limited to torsional potentials, or residue pair potentials, or residue entropy potentials. Such additional scoring functions can be used alone, or as functions for processing the library after it is scored initially. For example, a variety of functions derived from data on binding of peptides to MHC (Major Histocompabbility Complex) can be used to rescore a library in order to eliminate proteins containing sequences which can potentially bind to MHC, i.e. potentially immunogenic sequences.

In a preferred embodiment, a variety of filtering techniques can be done, including, but not limited to, DEE and its related counterparts. Additional filtering techniques include, but are not limited to branch-and-bound techniques for finding optimal sequences (Gordon and Mayo, Structure Fold. Des. 7:1089–98, 1999), and exhaustive enumeration of sequences.

As will be appreciated by those in the art, once an optimized sequence or set of sequences is generated, a variety of sequence space sampling methods can be done, either in addition to the preferred Monte Carlo methods, or instead of a Monte Carlo search. That is, once a sequence or set of sequences is generated, preferred methods utilize sampling techniques to allow the generation of additional, related sequences for testing.

These sampling methods can include the use of amino acid substitutions, insertions or deletions, or recombinations of one or more sequences. As outlined herein, a preferred embodiment utilizes a Monte Carlo search, which is a series of biased, systematic, or random jumps. However, there are other sampling techniques that can be used, including Boltzman sampling, genetic algorithm techniques and simulated annealing. In addition, for all the sampling techniques, the kinds of jumps allowed can be altered (e.g. random jumps to random residues, biased jumps (to or away from wild-type, for example), jumps to biased residues (to or away from similar residues, for example, etc.). Jumps where multiple residue positions are coupled (two residues always change together, or never change together), jumps where whole sets of residues change to other sequences (e.g., recombination). Similarly, for all the sampling techniques, the acceptance criteria of whether a sampling jump is accepted can be altered.

In addition, it should be noted that the preferred methods of the invention result in a rank ordered list of sequences; that is, the sequences are ranked on the basis of some objective criteria. However, as outlined herein, it is possible to create a set of non-ordered sequences, for example by generating a probability table directly (for example using SCMF analysis or sequence alignment techniques) that lists sequences without ranking them. The sampling techniques outlined herein can be used in either situation.

In a preferred embodiment, Boltzman sampling is done. As will be appreciated by those in the art, the temperature criteria for Boltzman sampling can be altered to allow broad searches at high temperature and narrow searches close to local optima at low temperatures (see e.g., Metropolis et al., J. Chem. Phys. 21:1087, 1953).

In a preferred embodiment, the sampling technique utilizes genetic algorithms, e.g., such as those described by Holland (Adaptation in Natural and Artifical Systems, 1975, Ann Arbor, U. Michigan Press). Genetic algorithm analysis generally takes generated sequences and recombines them computationally, similar to a nucleic acid recombination event, in a manner similar to "gene shuffling". Thus the "jumps" of genetic algorithm analysis generally are multiple position jumps. In addition, as outlined below, correlated multiple jumps may also be done. Such jumps can occur with different crossover positions and more than one recombination at a time, and can involve recombination of two or more sequences. Furthermore, deletions or insertions (random or biased) can be done. In addition, as outlined below, genetic algorithm analysis may also be used after the secondary library has been generated.

In a preferred embodiment, the sampling technique utilizes simulated annealing, e.g., such as described by Kirkpatrick et al. [Science, 220:671–680 (1983)]. Simulated annealing alters the cutoff for accepting good or bad jumps by altering the temperature. That is, the stringency of the cutoff is altered by altering the temperature. This allows broad searches at high temperature to new areas of sequence space, altering with narrow searches at low temperature to explore regions in detail.

In addition, as outlined below, these sampling methods can be used to further process a first set to generate additional sets of IbA proteins.

The computational processing results in a set of optimized IbA protein sequences. These optimized IbA protein sequences are generally significantly different from the wild-type IFN-β sequence from which the backbone was taken. That is, each optimized IbA protein sequence preferably comprises at least about 3–10% variant amino acids from the starting or wild type sequence, with at least about 10–15% being preferred, with at least about 15–20% changes being more preferred and at least 25% being particularly preferred.

In a preferred embodiment, the IbA proteins of the invention have 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 different residues from the human IFN-β sequence.

Thus, in the broadest sense, the present invention is directed to IbA proteins that have IFN-β activity. By "IFN-β activity" or "IbA" herein is meant that the IbA protein exhibits at least one, and preferably more, of the biological functions of an IFN-β, as defined below. In one embodiment, the biological function of an IbA protein is altered, preferably improved, over the corresponding biological activity of an IFN-β.

By "protein" herein is meant at least two covalently attached amino acids, which includes proteins, polypeptides, oligopeptides and peptides. The protein may be made up of naturally occurring amino acids and peptide bonds, or synthetic peptidomimetic structures, i.e., "analogs" such as peptoids [see Simon et al., Proc. Natl. Acd. Sci. U.S.A. 89(20:9367–71 (1992)], generally depending on the method of synthesis. Thus "amino acid", or "peptide residue", as used herein means both naturally occurring and synthetic amino acids. For example, homo-phenylalanine, citrulline, and noreleucine are considered amino acids for the purposes of the invention. "Amino acid" also includes amino acid residues such as proline and hydroxyproline. In addition, any amino acid representing a component of the IbA proteins can be replaced by the same amino acid but of the opposite chirality. Thus, any amino acid naturally occurring in the L-configuration (which may also be referred to as the R or S, depending upon the structure of the chemical entity) may be replaced with an amino acid of the same chemical structural type, but of the opposite chirality, generally referred to as the D- amino acid but which can additionally be referred to as the R- or the S-, depending upon its composition and chemical configuration. Such derivatives have the property of greatly increased stability, and therefore are advantageous in the formulation of compounds which may have longer in vivo half lives, when administered by oral, intravenous, intramuscular, intraperitoneal, topical, rectal, intraocular, or other routes. In the preferred embodiment, the amino acids are in the (S) or L-configuration. If non-naturally occurring side chains are used, non-amino acid substituents may be used, for example to prevent or retard in vivo degradations. Proteins including non-naturally occurring amino acids may be synthesized or in some cases, made recombinantly; see van Hest et al., FEBS Left 428:(1–2) 68–70 May 22, 1998 and Tang et al., Abstr. Pap Am. Chem. S218:U138-U138 Part 2 Aug. 22, 1999, both of which are expressly incorporated by reference herein.

Additionally, modified amino acids or chemical derivatives of amino acids of consensus or fragments of IbA proteins, according to the present invention may be provided, which polypeptides contain additional chemical moieties or modified amino acids not normally a part of the protein. Covalent and non-covalent modifications of the protein are thus included within the scope of the present invention.

Such modifications may be introduced into an IbA polypeptide by reacting targeted amino acid residues of the polypeptide with an organic derivatizing agent that is capable of reacting with selected side chains or terminal residues. The following examples of chemical derivatives are provided by way of illustration and not by way of limitation.

Aromatic amino acids may be replaced with D- or L-naphylalanine, D- or L-Phenylglycine, D- or L-2-thieneylalanine, D- or L-1-, 2-, 3- or 4-pyreneylalanine, D- or L-3-thieneylalanine, D- or L-(2-pyridinyl)-alanine, D- or L-(3-pyridinyl)-alanine, D- or L-(2-pyrazinyl)-alanine, D- or L-(4-isopropyl)-phenylglycine, D-(trifluoromethyl)-phenylglycine, D-(trifluoromethyl)-phenylalanine, D-p-fluorophenylalanine, D- or L-p-biphenylphenylalanine, D- or L-p-methoxybiphenylphenylalanine, D- or L-2-indole (alkyl)alanines, and D- or L-alkylainines where alkyl may be substituted or unsubstituted methyl, ethyl, propyl, hexyl, butyl, pentyl, isopropyl, iso-butyl, sec-isotyl, iso-pentyl, non-acidic amino acids, of C1–C20.

Acidic amino acids can be substituted with non-carboxylate amino acids while maintaining a negative charge, and derivatives or analogs thereof, such as the non-limiting examples of (phosphono)alanine, (phosphono) glycine, (phosphono)leucine, (phosphono)isoleucine, (phosphono)threonine, or (phosphono)senne; or sulfated (e.g., $—SO_3H$) threonine, serine, tyrosine.

Other substitutions may include unnatural hyroxylated amino acids that may be made by combining "alkyl" with any natural amino acid. The term "alkyl" as used herein refers to a branched or unbranched saturated hydrocarbon group of 1 to 24 carbon atoms, such as methyl, ethyl, n-propyl, isoptopyl, n-butyl, isobutyl, t-butyl, octyl, decyl, tetradecyl, hexadecyl, eicosyl, tetracisyl and the like. Preferred alkyl groups herein contain 1 to 12 carbon atoms. Also included within the definition of an alkyl group are cycloalkyl groups such as C5 and C6 rings, and heterocyclic rings with nitrogen, oxygen, sulfur or phosphorus. Alkyl also includes heteroalkyl, with heteroatoms of sulfur, oxygen, and nitrogen being preferred. Alkyl includes substituted alkyl groups. By "substituted alkyl group" herein is meant an alkyl group further comprising one or more substitution moieties. A preferred heteroalkyl group is an alkyl amine. By "alkyl amine" or grammatical equivalents herein is meant an alkyl group as defined above, substituted with an amine group at any position. In addition, the alkyl amine may have other substitution groups, as outlined above for alkyl group. The amine may be primary (—NH_2R), secondary (—NHR), or tertiary (—NR_3). Basic amino acids may be substituted with alkyl groups at any position of the naturally occurring amino acids lysine, arginine, ornithine, citrulline, or (guanidino)-acetic acid, or other (guanidino) alkyl-acetic acids, where "alkyl" is define as above. Nitrile derivatives (e.g., containing the CN-moiety in place of COOH) may also be substituted for asparagine or glutamine, and methionine sulfoxide may be substituted for methionine. Methods of preparation of such peptide derivatives are well known to one skilled in the art.

In addition, any amide linkage in any of the IbA polypeptides can be replaced by a ketomethylene moiety. Such derivatives are expected to have the property of increased stability to degradation by enzymes, and therefore possess advantages for the formulation of compounds which may have increased in vivo half lives, as administered by oral, intravenous, intramuscular, intraperitoneal, topical, rectal, intraocular, or other routes.

Additional amino acid modifications of amino acids of IbA polypeptides of the present invention may include the following: Cysteinyl residues may be reacted with alpha-haloacetates (and corresponding amines), such as 2-chloroacetic acid or chloroacetamide, to give carboxymethyl or carboxyamidomethyl derivatives. Cysteinyl residues may also be derivatized by reaction with compounds such as bromotrifluoroacetone, alpha-bromo-beta-(5-imidozoyl)propionic acid, chloroacetyl phosphate, N-alkylmaleimides, 3-nitro-2-pyridyl disulfide, methyl 2-pyridyl disulfide, p-chloromercuribenzoate, 2-chloromercuri-4-nitrophenol, or chloro-7-nitrobenzo-2-oxa-1,3-diazole.

Histidyl residues may be derivatzed by reaction with compounds such as diethylprocarbonate e.g., at pH 5.5–7.0 because this agent is relatively specific for the histidyl side chain, and para-bromophenacyl bromide may also be used; e.g., where the reaction is preferably performed in 0.1M sodium cacodylate at pH 6.0.

Lysinyl and amino terminal residues may be reacted with compounds such as succinic or other carboxylic acid anhydrides. Derivatization with these agents is expected to have the effect of reversing the charge of the lysinyl residues. Other suitable reagents for derivatizing alpha-amino-containing residues include compounds such as imidoesters/ e.g., as methyl picolinimidate; pyridoxal phosphate; pyridoxal; chloroborohydride; trinitrobenzenesulfonic acid; O-methylisourea; 2,4 pentanedione; and transaminase-catalyzed reaction with glyoxylate.

Arginyl residues may be modified by reaction with one or several conventional reagents, among them phenylglyoxal, 2,3-butanedione, 1,2-cyclohexanedione, and ninhydrin according to known method steps. Derivatization of arginine residues requires that the reaction be performed in alkaline conditions because of the high pKa of the guanidine functional group. Furthermore, these reagents may react with the groups of lysine as well as the arginine epsilon-amino group.

The specific modification of tyrosyl residues per se is well-known, such as for introducing spectral labels into tyrosyl residues by reaction with aromatic diazonium compounds or tetranitromethane. N-acetylimidizol and tetranitromethane may be used to form O-acetyl tyrosyl species and 3-nitro derivatives, respectively.

Carboxyl side groups (aspartyl or glutamyl) may be selectively modified by reaction with carbodiimides (R'-N-C-N-R') such as 1-cyclohexyl-3-(2-morpholinyl-(4-ethyl) carbodiimide or 1-ethyl-3-(4-azonia-4,4-dimethylpentyl) carbodiimide. Furthermore aspartyl and glutamyl residues may be converted to asparaginyl and glutaminyl residues by reaction with ammonium ions.

Glutaminyl and asparaginyl residues may be frequently deamidated to the corresponding glutamyl and aspartyl residues. Alternatively, these residues may be deamidated under mildly acidic conditions. Either form of these residues falls within the scope of the present invention.

The IFN-β may be from any number of organisms, with IFN-β s from mammals being particularly preferred. Suitable mammals include, but are not limited to, rodents (rats, mice, hamsters, guinea pigs, etc.), primates, farm animals (including sheep, goats, pigs, cows, horses, etc) and in the most preferred embodiment, from humans (this is sometimes referred to herein as hIFN-β, the sequence of which is depicted in FIG. 1). As will be appreciated by those in the art, IFN-β s based on IFN-β s from mammals other than humans may find use in animal models of human disease. The GenBank accession numbers for a variety of mammalian IFN-β species is as follows: bovine 69689, 124465 (IFN-β-1 precursor), 69688, 124467 (IFN-β-3 precursor), 69687, 124466 (IFN-β-2 precursor); dog 442673; sheep 310382; cat CAA69853, 1754718; pig 2411469, 164517; mouse 69686, 6754304, 51551, 124470, 494203; rat 7438651, 2497434, 1616939; Macaca fascicularis 3766295; horse 69685, 124468, 164229; human 69684, 124469, 4504603, 3318961, 3318960.

The IbA proteins of the invention exhibit at least one biological function of an IFN-β. By "interferon-beta" or "IFN-β" herein is meant a wild type IFN-β or an allelic variant thereof. Thus, IFN-β refers to all forms of IFN-β that are active in accepted IFN-β assays.

The IbA proteins of the invention exhibit at least one biological function of an IFN-β. By "biological function" or "biological property" herein is meant any one of the properties or functions of an IFN-β, including, but not limited to, the ability to effect cellular growth, in particular inhibition of cell proliferation; the ability to effect cellular differentiation, in particular induction of cell differentiation; the ability to induce changes in cell morphology; the ability to modulate the immune system; the ability to enhance histocompafibility antigen expression; the ability to stimulate immunoglobulin-Fc receptor expression on macrophages; the ability to induce antibody production in B lymphocytes, the ability to activate natural killer cells; the ability to bind to an IFN receptor; the ability to bind to a cell comprising an IFN receptor, the ability to treat multiple sclerosis; the ability to treat idiopathic pulmonary fibrosis; the ability to treat inflammatory diseases; the ability to treat viral diseases, including treatment of infections caused by papilloma viruses, such as genital warts and condylomata of the uterine cervix; hepatitis viruses, such as acute/chronic hepatitis B and non-A, non-B hepatitis (hepatitis C); herpes viruses, such as herpes genitalis, herpes zoster, herpes keratitis, and herpes simplex; viral encephalitis; cytomegalovirus pneumonia; and prophylaxis of rhinovirus; the ability to treat cancer, including treatment of several malignant diseases such as osteosarcoma, basal cell carcinoma, cervical dysplasia, glioma, acute myeloid leukemia, multiple myeloma, Hodgkin's disease, melanoma, renal cancer, liver cancer, and breast cancer.

All of these IbA proteins will exhibit at least 50% of the receptor binding or biological activity as the wild type IFN-β. More preferred are IbA proteins that exhibit at least 75%, even more preferred are IbA proteins that exhibit at least 90%, and most preferred are IbA proteins that exhibit more than 100% of the receptor binding or biological activity as the wild type IFN-β. Biological assays, receptor binding assays, anti-viral and anti-proliferation assays are described in U.S. patents 4,450,103; 4,518,584; 4,588,585; 4,737,462; 4,738,844; 4,738,845; 4,753,795; 4,769,233; 4,793,995; 4,914,033; 4,959,314; 5,183,746; 5,376,567; 5,545,723; 5,730,969; 5,814,485; 5,869,603 and in e.g., Anderson et al., J. Biol. Chem. 257(19):11301–4 (1982); Herberman et al., Nature 277(5693):221–3 (1979); Williams et al., Nature 282(5739):582–6 (1979); Branca and Baglioni, Nature 294(5843):768–70 (1981); Proc. Natl. Acad. Sci. U.S.A. 81(18):5662–6 (1984); Fellous et al., Proc. Natl. Acad. Sci. U.S.A. 79(10):3082–6 (1982); and Runkel et al., J. Biol. Chem. 273(14):8003–8 (1998), all of which are expressly incorporated by reference.

In one embodiment, at least one biological property of the IbA protein is altered when compared to the same property of IFN-β. As outlined above, the invention provides IbA nucleic acids encoding IbA polypeptides. The IbA polypeptide preferably has at least one property, which is substantially different from the same property of the corresponding naturally occurring IFN-β polypeptide. The property of the IbA polypeptide is the result the PDA analysis of the present invention.

The term "altered property" or grammatical equivalents thereof in the context of a polypeptide, as used herein, refer to any characteristic or attribute of a polypeptide that can be selected or detected and compared to the corresponding property of a naturally occurring protein. These properties include, but are not limited to oxidative stability, substrate specificity, substrate binding or catalytic activity, thermal stability, alkaline stability, pH activity profile, resistance to proteolytic degradation, Km, kcat, Km/kcat ratio, kinetic association (K_{on}) and dissociation (K_{off}) rate, protein folding, inducing an immune response, ability to bind to a ligand, ability to bind to a receptor, ability to be secreted, ability to be displayed on the surface of a cell, ability to oligomerize, ability to signal, ability to stimulate cell proliferation, ability to inhibit cell proliferation, ability to induce apoptosis, ability to be modified by phosphorylabon or glycosylation, ability to treat disease.

Unless otherwise specified, a substantial change in any of the above-listed properties, when comparing the property of an IbA polypeptide to the property of a naturally occurring IFN-β protein is preferably at least a 20%, more preferably, 50%, more preferably at least a 2-fold increase or decrease.

A change in oxidative stability is evidenced by at least about 20%, more preferably at least 50% increase of activity of an IbA protein when exposed to various oxidizing conditions as compared to that of IFN-β. Oxidative stability is measured by known procedures.

A change in alkaline stability is evidenced by at least about a 5% or greater increase or decrease (preferably increase) in the half life of the activity of an IbA protein when exposed to increasing or decreasing pH conditions as compared to that of IFN-β. Generally, alkaline stability is measured by known procedures.

A change in thermal stability is evidenced by at least about a 5% or greater increase or decrease (preferably increase) in the half life of the activity of an IbA protein when exposed to a relatively high temperature and neutral pH as compared to that of IFN-β. Generally, thermal stability is measured by known procedures.

Similarly, IbA proteins, for example are experimentally tested and validated in in vivo and in in vitro assays. Suitable assays include, but are not limited to, e.g., examining their binding affinity to natural occurring or variant receptors and to high affinity agonists and/or antagonists. In addition to cell-free biochemical affinity tests, quantitative comparison are made comparing kinetic and equilibrium binding constants for the natural receptor to the naturally occurring IFN-β and to the IbA proteins. The kinetic association rate (K_{on}) and dissociation rate (K_{off}), and the equilibrium binding constants (K_d) can be determined using surface plasmon resonance on a BIAcore instrument following the standard procedure in the literature [Pearce et al., Biochemistry 38:81–89 (1999)]. Comparing the binding constant between a natural receptor and its corresponding naturally occurring IFN-β with the binding constant of a natural occurring receptor and an IbA protein are made in order to evaluate the sensitivity and specificity of the IbA protein. Preferably, binding affinity of the IbA protein to natural receptors and agonists increases relative to the naturally occurring IFN-β, while antagonist affinity decreases. IbA proteins with higher affinity to antagonists relative to the IFN-β may also be generated by the methods of the invention.

As described above, one biological function of an IbA protein is the ability of the IbA protein to bind to cells comprising an interferon receptor.

In a preferred embodiment, the assay system used to determine IbA is an in vitro system using cells that either express endogenous interferon receptors or cells stably transfected with the gene encoding the human interferon receptor. In this system, cell proliferation is measured as a function of BrdU incorporation, which is incorporated into the nucleic acid of proliferating cells. A decrease above background of at least about 10%, with at least about 20% being preferred, with at least about 30% being more preferred and at least about 50%, 75% and 90% being especially preferred is an indication of IbA.

In a preferred embodiment, the antigenic profile in the host animal of the IbA protein is similar, and preferably identical, to the antigenic profile of the host IFN-β; that is, the IbA protein does not significantly stimulate the host organism (e.g. the patient) to an immune response; that is, any immune response is not clinically relevant and there is no allergic response or neutralization of the protein by an antibody. That is, in a preferred embodiment, the IbA protein does not contain additional or different epitopes from the IFN-β. By "epitope" or "determinant" herein is meant a portion of a protein which will generate and/or bind an antibody. Thus, in most instances, no significant amount of antibodies are generated to a IbA protein. In general, this is accomplished by not significantly altering surface residues, as outlined below nor by adding any amino acid residues on the surface which can become glycosylated, as novel glycosylation can result in an immune response.

The IbA proteins and nucleic acids of the invention are distinguishable from naturally occurring IFN-ps. By "naturally occurring" or "wild type" or grammatical equivalents, herein is meant an amino acid sequence or a nucleotide sequence that is found in nature and includes allelic variations; that is, an amino acid sequence or a nucleotide sequence that usually has not been intentionally modified. Accordingly, by "non-naturally occurring" or "synthetic" or "recombinant" or grammatical equivalents thereof, herein is meant an amino acid sequence or a nucleotide sequence that is not found in nature; that is, an amino acid sequence or a nucleotide sequence that usually has been intentionally modified. It is understood that once a recombinant nucleic acid is made and reintroduced into a host cell or organism, it will replicate non-recombinantly, i.e., using the in vivo cellular machinery of the host cell rather than in vitro manipulations, however, such nucleic acids, once produced recombinantly, although subsequently replicated non-recombinantly, are still considered recombinant for the purpose of the invention. Representative amino acid and nucleotide sequences of a naturally occurring human IFN-β are shown in FIG. 1. It should be noted that unless otherwise stated, all positional numbering of IbA proteins and IbA nucleic acids is based on these sequences. That is, as will be appreciated by those in the art, an alignment of IFN-β proteins and IbA proteins can be done using standard programs, as is outlined below, with the identification of "equivalent" positions between the two proteins. Thus, the IbA proteins and nucleic acids of the invention are non-naturally occurring; that is, they do not exist in nature.

Thus, in a preferred embodiment, the IbA protein has an amino acid sequence that differs from a wild-type IFN-β sequence by at least 3% of the residues. That is, the IbA proteins of the invention are less than about 97% identical to an IFN-β amino acid sequence. Accordingly, a protein is an "IbA protein" if the overall homology of the protein sequence to the amino acid sequence shown in FIG. 1A or FIG. 1B (SEQ ID NO:1) is preferably less than about 97%, more preferably less than about 95%, even more preferably less than about 90% and most preferably less than 85%. In some embodiments the homology will be as low as about 75 to 80%. Stated differently, based on the human IFN-β sequence of 166 residues (see FIG. 1A) (SEQ ID NO:1), IbA proteins have at least about 5 residues that differ from the human IFN-β sequence (3%), with IbA proteins having from 5 residues to upwards of 62 residues being different from the human IFN-β sequence. Preferred IbA proteins have 5–30 different residues with from about 5 to about 15 being particularly preferred (that is, 3–9% of the protein is not identical to human IFN-β).

In another preferred embodiment, IbA proteins have 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 different residues from the human IFN-β sequence.

Homology in this context means sequence similarity or identity, with identity being preferred. As is known in the art, a number of different programs can be used to identify whether a protein (or nucleic acid as discussed below) has sequence identity or similarity to a known sequence. Sequence identity and/or similarity is determined using standard techniques known in the art, including, but not limited to, the local sequence identity algorithm of Smith & Waterman, Adv. Appl. Math., 2:482 (1981), by the sequence identity alignment algorithm of Needleman & Wunsch, J. Mol. Biol., 48:443 (1970), by the search for similarity method of Pearson & Lipman, Proc. Natl. Acad. Sci. U.S.A., 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Drive, Madison, Wis.), the Best Fit sequence program described by Devereux et al., Nucl. Acid Res., 12:387–395 (1984), preferably using the default settings, or by inspection. Preferably, percent identity is calculated by FastDB based upon the following parameters: mismatch penalty of 1; gap penalty of 1; gap size penalty of 0.33; and joining penalty of 30, "Current Methods in Sequence Comparison and Analysis," Macromolecule Sequencing and Synthesis, Selected Methods and Applications, pp 127–149 (1988), Alan R. Liss, Inc.

An example of a useful algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Dooliftle, J. Mol. Evol. 35:351–360 (1987); the method is similar to that described by Higgins & Sharp CABIOS 5:151–153 (1989). Useful PILEUP parameters including a default gap weight of 3.00, a default gap length weight of 0.10, and weighted end gaps.

Another example of a useful algorithm is the BLAST algorithm, described in: Altschul et al., J. Mol. Biol. 215, 403–410, (1990); Altschul et al., Nucleic Acids Res. 25:3389–3402 (1997); and Karlin et al., Proc. Natl. Acad. Sci. U.S.A. 90:5873–5787 (1993). A particularly useful BLAST program is the WU-BLAST-2 program which was obtained from Altschul et al., Methods in Enzymology, 266:460–480 (1996); http:flblast.wustl/edu/blastt README.html]. WU-BLAST-2 uses several search parameters, most of which are set to the default values. The adjustable parameters are set with the following values: overlap span=1, overlap fraction=0.125, word threshold (T)=11. The HSP S and HSP S2 parameters are dynamic values and are established by the program itself depending upon the composition of the particular sequence and composition of the particular database against which the sequence of interest is being searched; however, the values may be adjusted to increase sensitivity.

An additional useful algorithm is gapped BLAST as reported by Altschul et al., Nucl. Acids Res., 25:3389–3402. Gapped BLAST uses BLOSUM-62 substitution scores; threshold T parameter set to 9; the two-hit method to trigger ungapped extensions; charges gap lengths of k a cost of 10+k; X_u set to 16, and X_g set to 40 for database search stage and to 67 for the output stage of the algorithms. Gapped alignments are triggered by a score corresponding to ~22 bits.

A % amino acid sequence identity value is determined by the number of matching identical residues divided by the total number of residues of the "longer" sequence in the aligned region. The "longer" sequence is the one having the most actual residues in the aligned region (gaps introduced by WU-Blast-2 to maximize the alignment score are ignored).

In a similar manner, "percent (%) nucleic acid sequence identity" with respect to the coding sequence of the polypeptides identified herein is defined as the percentage of nucleotide residues in a candidate sequence that are identical with the nucleofide residues in the coding sequence of the cell cycle protein. A preferred method utilizes the BLASTN module of WU-BLAST-2 set to the default parameters, with overlap span and overlap fraction set to 1 and 0.125, respectively.

The alignment may include the introduction of gaps in the sequences to be aligned. In addition, for sequences which contain either more or fewer amino acids than the protein encoded by the sequence of FIG. 1, it is understood that in one embodiment, the percentage of sequence identity will be determined based on the number of identical amino acids in relation to the total number of amino acids. Thus, for example, sequence identity of sequences shorter than that shown in FIG. 1, as discussed below, will be determined using the number of amino acids in the shorter sequence, in one embodiment. In percent identity calculations relative weight is not assigned to various manifestations of sequence variation, such as, insertions, deletions, substitutions, etc.

In one embodiment, only identities are scored positively (+1) and all forms of sequence variation including gaps are assigned a value of "0", which obviates the need for a weighted scale or parameters as described below for sequence similarity calculations. Percent sequence identity can be calculated, for example, by dividing the number of matching identical residues by the total number of residues of the "shorter" sequence in the aligned region and multiplying by 100. The "longer" sequence is the one having the most actual residues in the aligned region.

Thus, IbA proteins of the present invention may be shorter or longer than the amino acid sequence shown in FIG. 1A (SEQ ID NO:1). Thus, in a preferred embodiment, included within the definition of IbA proteins are portions or fragments of the sequences depicted herein. Fragments of IbA proteins are considered IbA proteins if a) they share at least one antigenic epitope; b) have at least the indicated homology; c) and preferably have IbA biological activity as defined herein.

In a preferred embodiment, as is more fully outlined below, the IbA proteins include further amino acid variations, as compared to a wild type IFN-β, than those outlined herein. In addition, as outlined herein, any of the variations depicted herein may be combined in any way to form additional novel IbA proteins.

In addition, IbA proteins can be made that are longer than those depicted in the figures, for example, by the addition of epitope or purification tags, as outlined herein, the addition of other fusion sequences, etc. For example, the IbA proteins of the invention may be fused to other therapeutic proteins such as IL-11 or to other proteins such as Fc or serum albumin for pharmacokinetic purposes. See for example U.S. Pat. No. 5,766,883 and 5,876,969, both of which are expressly incorporated by reference.

In a preferred embodiment, the IbA proteins comprise variable residues in core residues.

Human IFN-β core residues are as follows: positions 1, 6, 10, 13, 14, 15, 17, 18, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 (see FIG. 3). Accordingly, in a preferred embodiment, IbA proteins have variable positions selected from these positions.

The structure of human IFN-β as reported by Karpasus et al. (supra) indicated that IFN-β forms a dimer consisting of an A-chain and a B-chain.

Thus, in one embodiment, variable residues for the A-chain are as follows: positions 1, 6, 10, 13, 14, 17, 18, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 (see FIG. 3). Accordingly, in a preferred embodiment, IbA proteins have variable positions selected from these positions.

Thus, in another embodiment, variable residues for the B-chain are as follows: positions 1, 6, 10, 13, 14, 15, 17, 18, 21, 38, 50, 55, 56, 58, 59, 61, 62, 63, 66, 69, 70, 72, 74, 76, 77, 81, 84, 87, 90, 91, 94, 95, 98, 102, 114, 115, 118, 122, 125, 126, 129, 130, 132, 133, 136, 138, 139, 142, 143, 144, 146, 147, 150, 151, 153, 154, 157, 159, 160, 161, 163, and 164 (see FIG. 3). Accordingly, in a preferred embodiment, IbA proteins have variable positions selected from these positions.

In a preferred embodiment, IbA proteins have variable positions selected solely from core residues of human IFN-β. Alternatively, at least a majority (51%) of the variable positions are selected from core residues, with at least about 75% of the variable positions being preferably selected from core residue positions, and at least about 90% of the variable positions being particularly preferred. A specifically preferred embodiment has only core variable positions altered as compared to human IFN-β.

Particularly preferred embodiments where IbA proteins have variable core positions as compared to human IFN-β are shown in the Figures.

In one embodiment, the variable core positions are altered to any of the other 19 amino acids. In a preferred embodiment, the variable core residues are chosen from Ala, Val, Phe, Ile, Leu, Tyr, Trp and Met. In another preferred embodiment, the variable core residues are chosen from Ala, Val, Leu, Ile, Phe, Tyr, and Trp. In another preferred embodiment, the variable core residues are chosen from Ala, Val, leu, Ile, and Gly. In another preferred embodiment, the variable core residues are chosen from Ala, Gly, Ser, Thr, Glu, Asp, Gln, Asn, and Cys.

In a preferred embodiment, the IbA protein of the invention has a sequence that differs from a wild-type human IFN-β protein in at least one amino acid position selected from positions 6, 13, 17, 21, 56, 30 59, 61, 62, 63, 66, 69, 84, 87, 91, 98, 102, 114, 118, 122, 129, 146, 150, 154, 157, 160, and 161; see also FIG. 3, which outlines sets of amino acid positions.

Preferred amino acids for each position, including the human IFN-β residues, are shown in FIGS. 4–16 (SEQ ID NOS:4–24). Thus, for example, for the A-chain of an IbA protein, at position 13, preferred amino acids are Phe, Tyr, Glu, and Ala; at position 17, a preferred amino acid is Asp; at position 69, a preferred amino acid is Val; at position 84 a preferred amino acid is lie; at position 87, a preferred amino acid is Phe; at position 91, a preferred amino acid is lie; at position 98, a preferred amino acid is Phe; at position 118, preferred amino acids are Ala, Val, and Cys; at position 122, preferred amino acids are Ile and Val; at position 146, a preferred amino acid is Ile; at position 157, a preferred amino acid is Leu; and at position 161, preferred amino acids are Ala and Cys.

For the B-chain of an IbA protein, at position13, preferred amino acids are Leu and Glu; at position 17, preferred amino acid are Ala and Thr; at position 56, a preferred amino acid is Leu; at position 63, a preferred amino acid is Phe; at position 84 a preferred amino acid is lie; at position 87, a preferred amino acid is Phe; at position 91, a preferred amino acid is Ile; at position 114, preferred amino acids are Phe and Leu; at position 118, preferred amino acids are Leu and Glu; at position 122, preferred amino acids are Ile and Phe; and at position 161, preferred amino acids are Ala and Glu. Preferred changes are as follows: L6A; L6F; S13F; S13Y; S13L; S131; S13A; S13G; S13G; S13T; S13C; S13E; C17A; C17L; C17V; C17D; C17T; C171; C17E; C17S; C17G; L211; L21V; L21A; L21Y; L21F; A56L; 159V; 159A; 159L; M621; M62V; M62L; L63A; L63F; L63Y; I66L; I66V; I66A; I69V; I69L; I69A; V84I; V84L; V84A; L87F; L87I; L87Y; L87V; L87A; L87W; V91I; V91A;

V91L; V91F; V91Y; V98A; L98F; G114F; G114L; S118A; S118V; S118C; S118L; S118E; L122I; L122V; L122A; L122F; L122Y; L122W; I129V; I129L; I129A; V146I; V146A; I150V; I150A; I150L; I150F; F154L; F154Y; F157V; I157V; I157L; I157A; L160I; L160V; L160A; L160F; L160Y; T115A; T161V; T161I; T161D; T161C; T161E; and T161G. These may be done either individually or in combination, with any combination being possible. However, as outlined herein, preferred embodiments utilize at least five, and preferably more, variable positions in each IbA protein.

Particularly preferred sequences for IbA proteins are selected from the group consisting of: [V841 and L87F (FIG. 4B and FIG. 10B) (SEQ ID NOS:4,18)]; [V84I, V91I, L98F, L122I, and I157L (see FIG. 5B) (SEQ ID NO:5)]; [S13F, I69V, V84I, V91I, L98F, S118A, L122I, V146I, I157L, and T161A (see FIG. 6B) (SEQ ID NO:6)]; [S13Y, I69V, V84I, V91I, L98F, S118V, L122V, V146I, I157L, and T161A (see FIG. 6C) (SEQ ID NO:7)]; [S13F, V84I, V91I, L98F, S118A, L122I, I157L, and T161A(see FIG. 6D) (SEQ ID NO:8)]; [S13F, C17D, I69V, V84I, V91I, L98F, S118A, L122I, VI146I, I157L, and T161A (see FIG. 7B) (SEQ ID NO:9)]; [S13Y, C17D, 169V, V841, V91I, L98F, S118V, L122A, V146I, I157L, and T161A (see FIG. 7C) (SEQ ID NO:10)]; [S13F, C17D, V84I, V91I, L98F, S118A, L122I, I157L, and T161A (see FIG. 70) (SEQ ID NO:11)]; [S13E, C17D, V84I, V91I, S118C, V146I, and T161C (see FIG. 8B) (SEQ ID NO:12)]; [S13A, V84I, V91I, S118C, V146I, I157L,·and T161C (see FIG. 8C) (SEQ ID NO:13)]; [S13E, C17D, V84I, V91I, S118C, and T161C (see FIG. 8D)(SEQ ID NO:14)]; [S13E, C17D, 169V, V84I, V91I, S118A, L122I, V146I, I157L, and T161(see FIG. 9B) (SEQ ID NO:15)]; [S13E, C17D, V84I, V91I, S118A, V146I, and 1157L (see FIG. 9C) (SEQ ID NO:16)]; [S13E, C17D, V84I, V91I, S118A, L122I, I157L, and T161A (see FIG. 9D) (SEQ ID NO:17)]; [A56L, L63F, V84I, L87F, V91I, and L122F (see FIG. 11B) (SEQ ID NO:19)]; [S13L, A56L, V84I, V91I, G114F, S118L, L122I, and T161A (see FIG. 12B) (SEQ ID NO:20)]; [S13L, C17A, A56L, V84I, L87F, V91L, G114F, S118L, L122I, and T161E (see FIG. 13B) (SEQ ID NO:21)]; [S13E, A56L, V84I, V91I, G114L, S118E, and T161E (see FIG. 14B) (SEQ ID NO:22)]; [C17T, A56L, V84I, V84I, V91I, S118E, G114L, S118E, and T161E (see FIG. 15B) (SEQ ID NO:23)]; and [C17T, A56L, V84I, V91I, S118E, and T161E (see FIG. 16B) (SEQ ID NO:24)].

Particularly preferred sequences for the A-chain of an IbA protein are selected from the group consisting of: [V84I and L87F (FIG. 4B) (SEQ ID NO:4)]; [V84I, V91I, L98F, L122I, and I157L (see FIG. 5B) (SEQ ID NO:5)]; [S13F, I69V, V84I, V91I, L98F, S118A, L122I, V146I, I157L, and T161A (see FIG. 6B)(SEQ ID NO:6)]; [S13Y, I69V, V84I, V91I, L98F, S118V, L122V, V146I, I157L, and T161A (see FIG. 6C) (SEQ ID NO:7)]; [S13F, V84I, V91I, L98F, S118A, L122I, I157L, and T161A (see FIG. 6D) (SEQ ID NO:8)]; [S13F, C17D, I69V, V84I, V91I, L98F, S118A, L122I, V146I, I157L, and T161A (see FIG. 7B) (SEQ ID NO:9)]; [S13Y, C17D, I69V, V84I, V91I, L98F, S118V, L122V, V146I, I157L, and T161A (see FIG. 7C) (SEQ ID NO:10)]; [S13F, C17D, V84I, V91I, L98F, S118A, L122I, I157L, and T161A (see FIG. 7D) (SEQ ID NO:11)]; [S13E, C17D, V84I, V91I, S118C, V146I, and T161C (see FIG. 8B) (SEQ ID NO:12)]; [S13A, V84I, V91I, S118C, V146I, I157L, and T161C (see FIG. 8C) (SEQ ID NO:13)]; [S13E, C17D, V84I, V91I, S118C, and T161C (see FIG. 8D) (SEQ ID NO:14)]; [S13E, C17D, I69V, V84I, V91I, S118A, L122I, V146I, I157L, and T161A (see FIG. 9B)(SEQ ID NO:15)]; [S13E, C17D, V84I, V91I, S118A, V146I, and I157L (see FIG. 9C) (SEQ ID NO:16)]; and [S13E, C17D, V84I, V9I, S118A, L122I, I157L, and T161A (see FIG. 9D) (SEQ ID NO:17)].

Particularly preferred sequences for the B-chain of an IbA protein are selected from the group consisting of: [V84I and L87F (FIG. 10B) (SEQ ID NO:18)]; [A56L, L63F, V84I, L87F, V91I, and L122F (see FIG. 11B) (SEQ ID NO:19)]; [S13L, A56L, V84I, V91I, G114F, S118L, L122I, and (see FIG. 12B) (SEQ ID NO:20)]; [S13L, C17A, A56L, V84I, L87F, V91L, G114F, S118L, L122I, and T161E (see FIG. 13B) (SEQ ID NO:21)]; [S13E, A56L, V84I, V91I, G114L, S118E, and T161E (see FIG. 14B) (SEQ ID NO:22)]; [C17T, A56L, V84I, V91I, G114L, S118E, and T161E (see FIG 15B) (SEQ ID NO:23)]; and [C17T, A56L, V84I, V91I, S118E, and T161E (see FIG. 16B) (SEQ ID NO:24)].

In a preferred embodiment, the IbA proteins of the invention are human IFN-β conformers. By "conformer" herein is meant a protein that has a protein backbone 3D structure that is virtually the same but has significant differences in the amino acid side chains. That is, the IbA proteins of the invention define a conformer set, wherein all of the proteins of the set share a backbone structure and yet have sequences that differ by at least 3–5%. The three dimensional backbone structure of an IbA protein thus substantially corresponds to the three dimensional backbone structure of human IFN-β. "Backbone" in this context means the non-side chain atoms: the nitrogen, carbonyl carbon and oxygen, and the α-carbon, and the hydrogens attached to the nitrogen and α-carbon. To be considered a conformer, a protein must have backbone atoms that are no more than 2 Å from the human IFN-β structure, with no more than 1.5 Å being preferred, and no more than 1 Å being particularly preferred. In general, these distances may be determined in two ways. In one embodiment, each potential conformer is crystallized and its three dimensional structure determined. Alternatively, as the former is quite tedious, the sequence of each potential conformer is run in the PDA program to determine whether it is a conformer.

IbA proteins may also be identified as being encoded by IbA nucleic acids. In the case of the nucleic acid, the overall homology of the nucleic acid sequence is commensurate with amino acid homology but takes into account the degeneracy in the genetic code and codon bias of different organisms. Accordingly, the nucleic acid sequence homology may be either lower or higher than that of the protein sequence, with lower homology being preferred.

In a preferred embodiment, an IbA nucleic acid encodes an IbA protein. As will be appreciated by those in the art, due to the degeneracy of the genetic code, an extremely large number of nucleic acids may be made, all of which encode the IbA proteins of the present invention. Thus, having identified a particular amino acid sequence, those skilled in the art could make any number of different nucleic acids, by simply modifying the sequence of one or more codons in a way which does not change the amino acid sequence of the IbA.

In one embodiment, the nucleic acid homology is determined through hybridization studies. Thus, for example, nucleic acids which hybridize under high stringency to the nucleic acid sequence shown in FIG. 1 or its complement and encode a IbA protein is considered an IbA gene.

High stringency conditions are known in the art; see for example Maniatis et al., Molecular Cloning: A Laboratory Manual, 2d Edition, 1989, and Short Protocols in Molecular Biology, ed. Ausubel, et al., both of which are hereby incorporated by reference. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Techniques in Biochemistry and Molecular Biology—Hybridization with Nucleic Acid Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, stringent conditions are selected to be about 5–10° C. lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength, pH and nucleic acid concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T_m, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30° C. for short probes (e.g. 10 to 50 nucleotides) and at least about 60° C. for long probes (e.g. greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

In another embodiment, less stringent hybridization conditions are used; for example, moderate or low stringency conditions may be used, as are known in the art; see Maniatis and Ausubel, supra, and Tijssen, supra.

The IbA proteins and nucleic acids of the present invention are recombinant. As used herein, "nucleic acid" may refer to either DNA or RNA, or molecules which contain both deoxy- and ribonucleotides. The nucleic acids include genomic DNA, cDNA and oligonucleotides including sense and anti-sense nucleic acids. Such nucleic acids may also contain modifications in the ribose-phosphate backbone to increase stability and half life of such molecules in physiological environments.

The nucleic acid may be double stranded, single stranded, or contain portions of both double stranded or single stranded sequence. As will be appreciated by those in the art, the depiction of a single strand ("Watson") also defines the sequence of the other strand ("Crick"); thus the sequence depicted in FIG. 1 also includes the complement of the sequence. By the term "recombinant nucleic acid" herein is meant nucleic acid, originally formed in vitro, in general, by the manipulation of nucleic acid by endonucleases, in a form not normally found in nature. Thus an isolated IbA nucleic acid, in a linear form, or an expression vector formed in vitro by ligating DNA molecules that are not normally joined, are both considered recombinant for the purposes of this invention. It is understood that once a recombinant nucleic acid is made and reintroduced into a host cell or organism, it will replicate non-recombinantly, i.e. using the in vivo cellular machinery of the host cell rather than in vitro manipulations; however, such nucleic acids, once produced recombinantly, although subsequently replicated non-recombinantly, are still considered recombinant for the purposes of the invention.

Similarly, a "recombinant protein" is a protein made using recombinant techniques, i.e. through the expression of a recombinant nucleic acid as depicted above. A recombinant protein is distinguished from naturally occurring protein by at least one or more characteristics. For example, the protein may be isolated or purified away from some or all of the proteins and compounds with which it is normally associated in its wild type host, and thus may be substantially pure. For example, an isolated protein is unaccompanied by at least some of the material with which it is normally associated in its natural state, preferably constituting at least about 0.5%, more preferably at least about 5% by weight of the total protein in a given sample. A substantially pure protein comprises at least about 75% by weight of the total protein, with at least about 80% being preferred, and at least about 90% being particularly preferred. The definition includes the production of an IbA protein from one organism in a different organism or host cell. Alternatively, the protein may be made at a significantly higher concentration than is normally seen, through the use of an inducible promoter or high expression promoter, such that the protein is made at increased concentration levels. Furthermore, all of the IbA proteins outlined herein are in a form not normally found in nature, as they contain amino acid substitutions, insertions and deletions, with substitutions being preferred, as discussed below.

Also included within the definition of IbA proteins of the present invention are amino acid sequence variants of the IbA sequences outlined herein and shown in the Figures. That is, the IbA proteins may contain additional variable positions as compared to human IFN-β. These variants fall into one or more of three classes: substitutional, insertional or deletional variants. These variants ordinarily are prepared by site specific mutagenesis of nucleotdes in the DNA encoding an IbA protein, using cassette or PCR mutagenesis or other techniques well known in the art, to produce DNA encoding the variant, and thereafter expressing the DNA in recombinant cell culture as outlined above. However, variant IbA protein fragments having up to about 100–150 residues may be prepared by in vitro synthesis using established techniques. Amino acid sequence variants are characterized by the predetermined nature of the variation, a feature that sets them apart from naturally occurring allelic or interspecies variation of the IbA protein amino acid sequence. The variants typically exhibit the same qualitative biological activity as the naturally occurring analogue, although variants can also be selected which have modified characteristics as will be more fully outlined below.

While the site or region for introducing an amino acid sequence variation is predetermined, the mutation per se need not be predetermined. For example, in order to optimize the performance of a mutation at a given site, random mutagenesis may be conducted at the target codon or region and the expressed IbA variants screened for the optimal combination of desired activity. Techniques for making substitution mutations at predetermined sites in DNA having a known sequence are well known, for example, M13 primer mutagenesis and PCR mutagenesis. Screening of the mutants is done using assays of IbA protein activities.

Amino acid substitutions are typically of single residues; insertions usually will be on the order of from about 1 to 20 amino acids, although considerably larger insertions may be tolerated. Deletions range from about 1 to about 20 residues, although in some cases deletions may be much larger.

Substitutions, deletions, insertions or any combination thereof may be used to arrive at a final derivative. Generally these changes are done on a few amino acids to minimize the alteration of the molecule. However, larger changes may be tolerated in certain circumstances. When small alterations in the characteristics of the IbA protein are desired, substitutions are generally made in accordance with the following chart:

CHART I

Original Residue	Exemplary Substitutions
Ala	Ser
Arg	Lys
Asn	Gln, His
Asp	Glu
Cys	Ser, Ala
Gln	Asn
Glu	Asp
Gly	Pro
His	Asn, Gln
Ile	Leu, Val
Leu	Ile, Val
Lys	Arg, Gln, Glu
Met	Leu, Ile
Phe	Met, Leu, Tyr
Ser	Thr
Thr	Ser
Trp	Tyr
Tyr	Trp, Phe
Val	Ile, Leu

Substantial changes in function or immunological identity are made by selecting substitutions that are less conservative than those shown in Chart I. For example, substitutions may be made which more significantly affect: the structure of the polypeptide backbone in the area of the alteration, for example the alpha-helical or beta-sheet structure; the charge or hydrophobicity of the molecule at the target site; or the bulk of the side chain. The substitutions which in general are expected to produce the greatest changes in the polypeptide's properties are those in which (a) a hydrophilic residue, e.g. seryl or threonyl, is substituted for (or by) a hydrophobic residue, e.g. leucyl, isoleucyl, phenylalanyl, valyl or alanyl; (b) a cysteine or proline is substituted for (or by) any other residue; (c) a residue having an electropositive side chain, e.g. lysyl, arginyl, or histidyl, is substituted for (or by) an electronegative residue, e.g. glutamyl or aspartyl; or (d) a residue having a bulky side chain, e.g. phenylalanine, is substituted for (or by) one not having a side chain, e.g. glycine.

The variants typically exhibit the same qualitative biological activity and will elicit the same immune response as the original IbA protein, although variants also are selected to modify the characteristics of the IbA proteins as needed. Alternatively, the variant may be designed such that the biological activity of the IbA protein is altered. For example, glycosylation sites may be altered or removed. Similarly, the biological function may be altered; for example, in some instances it may be desirable to have more or less potent IFN-β activity.

The IbA proteins and nucleic acids of the invention can be made in a number of ways. Individual nucleic acids and proteins can be made as known in the art and outlined below. Alternatively, libraries of IbA proteins can be made for testing.

In a preferred embodiment, sets or libraries of IbA proteins are generated from a probability distribution table. As outlined herein, there are a variety of methods of generating a probability distribution table, including using PDA, sequence alignments, forcefield calculations such as SCMF calculations, etc. In addition, the probability distribution can be used to generate information entropy scores for each position, as a measure of the mutational frequency observed in the library.

In this embodiment, the frequency of each amino acid residue at each variable position in the list is identified. Frequencies can be thresholded, wherein any variant frequency lower than a cutoff is set to zero. This cutoff is preferably 1%, 2%, 5%, 10% or 20%, with 10% being particularly preferred. These frequencies are then built into the IbA library. That is, as above, these variable positions are collected and all possible combinations are generated, but the amino acid residues that "fill" the library are utilized on a frequency basis. Thus, in a non-frequency based library, a variable position that has 5 possible residues will have 20% of the proteins comprising that variable position with the first possible residue, 20% with the second, etc. However, in a frequency based library, a variable position that has 5 possible residues with frequencies of 10%, 15%, 25%, 30% and 20%, respectively, will have 10% of the proteins comprising that variable position with the first possible residue, 15% of the proteins with the second residue, 25% with the third, etc. As will be appreciated by those in the art, the actual frequency may depend on the method used to actually generate the proteins; for example, exact frequencies may be possible when the proteins are synthesized. However, when the frequency-based primer system outlined below is used, the actual frequencies at each position will vary, as outlined below.

As will be appreciated by those in the art and outlined herein, probability distribution tables can be generated in a variety of ways. In addition to the methods outlined herein, self-consistent mean field (SCMF) methods can be used in the direct generation of probability tables. SCMF is a deterministic computational method that uses a mean field description of rotamer interactions to calculate energies. A probability table generated in this way can be used to create libraries as described herein. SCMF can be used in three ways: the frequencies of amino acids and rotamers for each amino acid are listed at each position; the probabilities are determined directly from SCMF (see Delarue et la. Pac. Symp. Biocomput. 109–21 (1997), expressly incorporated by reference). In addition, highly variable positions and non-variable positions can be identified. Alternatively, another method is used to determine what sequence is jumped to during a search of sequence space; SCMF is used to obtain an accurate energy for that sequence; this energy is then used to rank it and create a rank-ordered list of sequences (similar to a Monte Carlo sequence list). A probability table showing the frequencies of amino acids at each position can then be calculated from this list (Koehl et al., J. Mol. Biol. 239:249 (1994); Koehl et al., Nat. Struc. Biol. 2:163 (1995); Koehl et al., Curr. Opin. Struct. Biol. 6:222 (1996); Koehl et al., J. Mol. Bio. 293:1183 (1999); Koehl et al., J. Mol. Biol. 293:1161 (1999); Lee J. Mol. Biol. 236:918 (1994); and Vasquez Biopolymers 36:53–70 (1995); all of which are expressly incorporated by reference. Similar methods include, but are not limited to, OPLS-AA (Jorgensen, et al., J. Am. Chem. Soc. (1996), v 118, pp 11225–11236; Jorgensen, W. L.; BOSS, Version 4.1; Yale University: New Haven, Conn. (1999)); OPLS (Jorgensen, et al., J. Am. Chem. Soc. (1988), v 110, pp 1657ff; Jorgensen, et al., J Am. Chem. Soc. (1990), v 112, pp 4768ff); UNRES (United Residue Forcefield; Liwo, et al., Protein Science (1993), v 2, pp1697–1714; Liwo, et al., Protein Science (1993), v 2, pp1715–1731; Liwo, et al., J. Comp. Chem. (1997), v 18, pp849–873; Liwo, et al., J. Comp. Chem. (1997), v 18, pp874–884; Liwo, et al., J. Comp. Chem. (1998), v 19, pp259–276; Forcefield for Protein Structure Prediction (Liwo, et al., Proc. Natl. Acad. Sci. USA (1999), v 96, pp5482–5485); ECEPP/3 (Liwo et al., J Protein Chem 1994 May;13(4):375–80); AMBER 1.1 force field (Weiner, et al., J. Am. Chem. Soc. v106, pp765–784); AMBER 3.0 force field (U.C. Singh et al., Proc. Natl. Acad. Sci. USA. 82:755–759); CHARMM and CHARMM22 (Brooks, et al., J. Comp. Chem. v4, pp 187–217); cvff3.0 (Dauber-Osguthorpe, et al., (1988) Proteins: Structure, Function and Genetics, v4, pp31–47); cff91 (Maple, et al., J. Comp. Chem. v 5, 162–182); also, the DISCOVER (cvff and cff91) and AMBER forcefields are used in the INSIGHT molecular modeling package (Biosym/MSI, San Diego Calif.) and HARMM is used in the QUANTA molecular modeling package (Biosym/MSI, San Diego Calif.).

In addition, as outlined herein, a preferred method of generating a probability distribution table is through the use of sequence alignment programs. In addition, the probability table can be obtained by a combination of sequence alignments and computational approaches. For example, one can add amino acids found in the alignment of homologous sequences to the result of the computation. Preferable one can add the wild type amino acid identity to the probability table if it is not found in the computation.

As will be appreciated, an IbA library created by recombining variable positions and/or residues at the variable position may not be in a rank-ordered list. In some embodiments, the entire list may just be made and tested. Alternatively, in a preferred embodiment, the IbA library is also in the form of a rank ordered list. This may be done for several reasons, including the size of the library is still too big to generate experimentally, or for predictive purposes. This may be done in several ways. In one embodiment, the library is ranked using the scoring functions of PDA to rank the library members. Alternatively, statistical methods could be used. For example, the library may be ranked by frequency score; that is, proteins containing the most of high frequency residues could be ranked higher, etc. This may be done by adding or multiplying the frequency at each variable position to generate a numerical score. Similarly, the library different positions could be weighted and then the proteins scored; for example, those containing certain residues could be arbitrarily ranked.

In a preferred embodiment, the different protein members of the IbA library may be chemically synthesized. This is particularly useful when the designed proteins are short, preferably less than 150 amino acids in length, with less than 100 amino acids being preferred, and less than 50 amino acids being particularly preferred, although as is known in the art, longer proteins can be made chemically or enzymatically. See for example Wilken et al, Curr. Opin. Biotechnol. 9:412–26 (1998), hereby expressly incorporated by reference.

In a preferred embodiment, particularly for longer proteins or proteins for which large samples are desired, the library sequences are used to create nucleic acids such as DNA which encode the member sequences and which can then be cloned into host cells, expressed and assayed, if desired. Thus, nucleic acids, and particularly DNA, can be made which encodes each member protein sequence. This is done using well known procedures. The choice of codons, suitable expression vectors and suitable host cells will vary depending on a number of factors, and can be easily optimized as needed.

In a preferred embodiment, multiple PCR reactions with pooled oligonucleotides is done, as is generally depicted in FIG. 17. In this embodiment, overlapping oligonucleotides are synthesized which correspond to the full length gene. Again, these oligonucleotides may represent all of the different amino acids at each variant position or subsets.

In a preferred embodiment, these oligonucleotides are pooled in equal proportions and multiple PCR reactions are performed to create full length sequences containing the combinations of mutations defined by the library. In addition, this may be done using error-prone PCR methods.

In a preferred embodiment, the different oligonucleotides are added in relative amounts corresponding to the probability distribution table. The multiple PCR reactions thus result in full length sequences with the desired combinations of mutations in the desired proportions.

The total number of oligonucleotides needed is a function of the number of positions being mutated and the number of mutations being considered at these positions: (number of oligos for constant positions)+M1+M2+M3+ . . . Mn=(total number of oligos required), where Mn is the number of mutations considered at position n in the sequence.

In a preferred embodiment, each overlapping oligonucleotide comprises only one position to be varied; in alternate embodiments, the variant positions are too close together to allow this and multiple variants per oligonucleotide are used to allow complete recombination of all the possibilities. That is, each oligo can contain the codon for a single position being mutated, or for more than one position being mutated. The multiple positions being mutated must be close in sequence to prevent the oligo length from being impractical. For multiple mutating positions on an oligonucleotide, particular combinations of mutations can be included or excluded in the library by including or excluding the oligonucleotide encoding that combination. For example, as discussed herein, there may be correlations between variable regions; that is, when position X is a certain residue, position Y must (or must not) be a particular residue. These sets of variable positions are sometimes referred to herein as a "cluster". When the clusters are comprised of residues close together, and thus can reside on one oligonucleotide primer, the clusters can be set to the "good" correlations, and eliminate the bad combinations that may decrease the effectiveness of the library. However, if the residues of the cluster are far apart in sequence, and thus will reside on different oligonucleotides for synthesis, it may be desirable to either set the residues to the "good" correlation, or eliminate them as variable residues entirely. In an alternative embodiment, the library may be generated in several steps, so that the cluster mutations only appear together. This procedure, i.e. the procedure of identifying mutation clusters and either placing them on the same oligonucleotides or eliminating them from the library or library generation in several steps preserving clusters, can considerably enrich the experimental library with properly folded protein. Identification of clusters can be carried out by a number of ways, e.g. by using known pattern recognition methods, comparisons of frequencies of occurence of mutations or by using energy analysis of the sequences to be experimentally generated (for example, if the energy of interaction is high, the positions are correlated). These correlations may be positional correlations (e.g. variable positions 1 and 2 always change together or never change together) or sequence correlations (e.g. if there is residue A at position 1, there is always residue B at position 2). See: Pattern discovery in Biomolecular Data: Tools, Techniques, and Applications; edited by Jason T. L. Wang, Bruce A. Shapiro, Dennis Shasha. New York: Oxford University, 1999; Andrews, Harry C. Introduction to mathematical techniques in pattern recognition; New York, Wiley-Interscience [1 972]; Applications of Pattern Recognition; Editor, K. S. Fu. Boca Raton, Fla. CRC Press, 1982; Genetic Algorithms for Pattern Recognition; edited by Sankar K. Pal, Paul P. Wang. Boca Raton: CRC Press, c1996; Pandya, Abhijit S., Pattern recognition with neural networks in C++/Abhijit S. Pandya, Robert B. Macy. Boca Raton, Fla.: CRC Press, 1996; Handbook of pattern recognition & computer vision I edited by C. H. Chen, L. F. Pau, P. S. P. Wang. 2nd ed. Singapore; River Edge, N.J.: World Scientific, c1999; Friedman, Introduction to Pattern Recognition: Statistical, Structural, Neural, and Fuzzy Logic Approaches; River Edge, N.J.: World Scientific, c1999, Series title: Series in machine perception and artificial intelligence; vol. 32; all of which are expressly incorporated by reference. In addition, programs used to search for consensus motifs can be used as well.

In addition, correlations and shuffling can be fixed or optimized by altering the design of the oligonucleotides; that is, by deciding where the oligonucleotides (primers) start and stop (e.g. where the sequences are "cut"). The start and stop sites of oligos can be set to maximize the number of clusters that appear in single oligonucleotides, thereby enriching the library with higher scoring sequences. Different oligonucleotide start and stop site options can be computationally modeled and ranked according to number of clusters that are represented on single oligos, or the percentage of the resulting sequences consistent with the predicted library of sequences.

The total number of oligonucleotides required increases when multiple mutable positions are encoded by a single oligonucleotide. The annealed regions are the ones that remain constant, i.e. have the sequence of the reference sequence.

Oligonucleotides with insertions or deletions of codons can be used to create a library expressing different length proteins. In particular computational sequence screening for insertions or deletions can result in secondary libraries defining different length proteins, which can be expressed by a library of pooled oligonucleotide of different lengths.

In a preferred embodiment, the IbA library is done by shuffling the family (e.g. a set of variants); that is, some set of the top sequences (if a rank-ordered list is used) can be shuffled, either with or without error-prone PCR. "Shuffling" in this context means a recombination of related sequences, generally in a random way. It can include "shuffling" as defined and exemplified in U.S. Pat. Nos. 5,830,721; 5,811,238; 5,605,793; 5,837,458 and PCT US/19256, all of which are expressly incorporated by reference in their entirety. This set of sequences can also be an artificial set; for example, from a probability table (for example generated using SCMF) or a Monte Carlo set. Similarly, the "family" can be the top 10 and the bottom 10 sequences, the top 100 sequence, etc. This may also be done using error-prone PCR.

Thus, in a preferred embodiment, in silico shuffling is done using the computational methods described herein. That is, starting with either two libraries or two sequences, random recombinations of the sequences can be generated and evaluated.

In a preferred embodiment, error-prone PCR is done to generate the IbA library. See U.S. Pat. Nos. 5,605,793, 5,811,238, and 5,830,721, all of which are hereby incorporated by reference. This can be done on the optimal sequence or on top members of the library, or some other artificial set or family. In this embodiment, the gene for the optimal sequence found in the computational screen of the primary library can be synthesized. Error prone PCR is then performed on the optimal sequence gene in the presence of oligonucleofides that code for the mutations at the variant positions of the library (bias oligonucleotides). The addition of the oligonucleotdes will create a bias favoring the incorporation of the mutations in the library. Alternatively, only oligonucleotdes for certain mutations may be used to bias the library.

In a preferred embodiment, gene shuffling with error prone PCR can be performed on the gene for the optimal sequence, in the presence of bias oligonucleotides, to create a DNA sequence library that reflects the proportion of the mutations found in the IbA library. The choice of the bias oligonucleotides can be done in a variety of ways; they can be chosen on the basis of their frequency, i.e. oligonucleotides encoding high mutational frequency positions can be used; alternatively, oligonucleotides containing the most variable positions can be used, such that the diversity is increased; if the secondary library is ranked, some number of top scoring positions can be used to generate bias oligonucleotides; random positions may be chosen; a few top scoring and a few low scoring ones may be chosen; etc. What is important is to generate new sequences based on preferred variable positions and sequences.

In a preferred embodiment, PCR using a wild type gene or other gene can be used, as is schematically depicted in FIG. **18**. In this embodiment, a starting gene is used; generally, although this is not required, the gene is usually the wild type gene. In some cases it may be the gene encoding the global optimized sequence, or any other sequence of the list, or a consensus sequence obtained e.g. from aligning homologous sequences from different organisms. In this embodiment, oligonucleotides are used that correspond to the variant positions and contain the different amino acids of the library. PCR is done using PCR primers at the termini, as is known in the art. This provides two benefits; the first is that this generally requires fewer oligonucleotides and can result in fewer errors. In addition, it has experimental advantages in that if the wild type gene is used, it need not be synthesized.

In addition, there are several other techniques that can be used, as exemplified in the figures, e.g. FIGS. **19–21**. In a preferred embodiment, ligation of PCR products is done.

In a preferred embodiment, a variety of additional steps may be done to the IbA library; for example, further computational processing can occur, different IbA libraries can be recombined, or cutoffs from different libraries can be combined. In a preferred embodiment, an IbA library may be computationally remanipulated to form an additional IbA library (sometimes referred to herein as "tertiary libraries"). For example, any of the IbA library sequences may be chosen for a second round of PDA, by freezing or fixing some or all of the changed positions in the first library. Alternatively, only changes seen in the last probability distribution table are allowed. Alternatively, the stringency of the probability table may be altered, either by increasing or decreasing the cutoff for inclusion. Similarly, the IbA library may be recombined experimentally after the first round; for example, the best gene/genes from the first screen may be taken and gene assembly redone (using techniques outlined below, multiple PCR, error prone PCR, shuffling, etc.). Alternatively, the fragments from one or more good gene(s) to change probabilities at some positions. This biases the search to an area of sequence space found in the first round of computational and experimental screening.

In a preferred embodiment, a tertiary library can be generated from combining different IbA libraries. For example, a probability distribution table from a first IbA library can be generated and recombined, either computationally or experimentally, as outlined herein. A PDA IbA library may be combined with a sequence alignment IbA library, and either recombined (again, computationally or experimentally) or just the cutoffs from each joined to make a new tertiary library. The top sequences from several libraries can be recombined. Sequences from the top of a library can be combined with sequences from the bottom of the library to more broadly sample sequence space, or only sequences distant from the top of the library can be combined. IbA libraries that analyzed different parts of a protein can be combined to a tertiary library that treats the combined parts of the protein.

In a preferred embodiment, a tertiary library can be generated using correlations in an IbA library. That is, a residue at a first variable position may be correlated to a residue at second variable position (or correlated to residues at additional positions as well). For example, two variable positions may sterically or electrostatically interact, such that if the first residue is X, the second residue must be Y. This may be either a positive or negative correlation.

Using the nucleic acids of the present invention which encode an IbA protein, a variety of expression vectors are made. The expression vectors may be either self-replicating extrachromosomal vectors or vectors which integrate into a host genome. Generally, these expression vectors include transcriptional and translational regulatory nucleic acid operably linked to the nucleic acid encoding the IbA protein. The term "control sequences" refers to DNA sequences necessary for the expression of an operably linked coding sequence in a particular host organism. The control sequences that are suitable for prokaryotes, for example, include a promoter, optionally an operator sequence, and a ribosome binding site. Eukaryotic cells are known to utilize promoters, polyadenylation signals, and enhancers.

Nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For example, DNA for a presequence or secretory leader is operably linked to DNA for a polypeptide if it is expressed as a preprotein that participates in the secretion of the polypeptide; a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the sequence; or a ribosome binding site is operably linked to a coding sequence if it is positioned so as to facilitate translation.

In a preferred embodiment, when the endogenous secretory sequence leads to a low level of secretion of the naturally occurring protein or of the IbA protein, a replacement of the naturally occurring secretory leader sequence is desired. In this embodiment, an unrelated secretory leader sequence is operably linked to an IbA encoding nucleic acid leading to increased protein secretion. Thus, any secretory leader sequence resulting in enhanced secretion of the IbA protein, when compared to the secretion of IFN-β and its secretory sequence, is desired. Suitable secretory leader sequences that lead to the secretion of a protein are know in the art.

In another preferred embodiment, a secretory leader sequence of a naturally occurring protein or a protein is removed by techniques known in the art and subsequent expression results in intracellular accumulation of the recombinant protein.

Generally, "operably linked" means that the DNA sequences being linked are contiguous, and, in the case of a secretory leader, contiguous and in reading phase. However, enhancers do not have to be contiguous. Linking is accomplished by ligation at convenient restriction sites. If such sites do not exist, the synthetic oligonucleotide adaptors or linkers are used in accordance with conventional practice. The transcriptional and translational regulatory nucleic acid will generally be appropriate to the host cell used to express the fusion protein; for example, transcriptional and translational regulatory nucleic acid sequences from Bacillus are preferably used to express the fusion protein in Bacillus. Numerous types of appropriate expression vectors, and suitable regulatory sequences are known in the art for a variety of host cells.

In general, the transcriptional and translational regulatory sequences may include, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, and enhancer or activator sequences. In a preferred embodiment, the regulatory sequences include a promoter and transcriptional start and stop sequences.

Promoter sequences encode either constitutive or inducible promoters. The promoters may be either naturally occurring promoters or hybrid promoters. Hybrid promoters, which combine elements of more than one promoter, are also known in the art, and are useful in the present invention. In a preferred embodiment, the promoters are strong promoters, allowing high expression in cells, particularly mammalian cells, such as the CMV promoter, particularly in combination with a Tet regulatory element.

In addition, the expression vector may comprise additional elements. For example, the expression vector may have two replication systems, thus allowing it to be maintained in two organisms, for example in mammalian or insect cells for expression and in a prokaryotic host for cloning and amplification. Furthermore, for integrating expression vectors, the expression vector contains at least one sequence homologous to the host cell genome, and preferably two homologous sequences which flank the expression construct. The integrating vector may be directed to a specific locus in the host cell by selecting the appropriate homologous sequence for inclusion in the vector. Constructs for integrating vectors are well known in the art.

In addition, in a preferred embodiment, the expression vector contains a selectable marker gene to allow the selection of transformed host cells. Selection genes are well known in the art and will vary with the host cell used.

A preferred expression vector system is a retroviral vector system such as is generally described in PCT/US97/01019 and PCT/US97/01048, both of which are hereby expressly incorporated by reference.

In a preferred embodiment, the expression vector comprises the components described above and a gene encoding an IbA protein. In this aspect, only one species of an IbA protein will be expressed in the cell comprising the expression vector. In one aspect of this embodiment, it is desired to express an optimized A-chain of IFN-β and an optimized B-chain of IFN-β within the same cell and thus, two expression vectors, one comprising a gene coding for an optimized A-chain of IFN-β, the other one comprising a gene coding for an optimized B-chain of IFN-β are introduced into the same host cell. This allows formation of a preferred IbA dimer.

In another aspect of this embodiment, an expression vector is constructed that comprises two IbA genes encoding two different IbA proteins. In this embodiment, one IbA gene encodes an optimized A chain of IFN-β and the second gene encodes an optimized B-chain of IFN-β. In one aspect of this embodiment, a polycistronic gene can be constructed as is known in the art for co-expression in a host cell.

As will be appreciated by those in the art, all combinations are possible and accordingly, as used herein, the combination of components, comprised by one or more vectors, which may be retroviral or not, is referred to herein as a "vector composition".

The IbA nucleic acids are introduced into the cells either alone or in combination with an expression vector. By "introduced into" or grammatical equivalents herein is meant that the nucleic acids enter the cells in a manner suitable for subsequent expression of the nucleic acid. The method of introduction is largely dictated by the targeted cell type, discussed below. Exemplary methods include $(Ca_3PO_4)_2$ precipitation, liposome fusion, lipofectin®, electroporation, viral infection, etc. The IbA nucleic acids may stably integrate into the genome of the host cell (for example, with retroviral introduction, outlined below), or may exist either transiently or stably in the cytoplasm (i.e. through the use of traditional plasmids, utilizing standard regulatory sequences, selection markers, etc.).

The IbA proteins of the present invention are produced by culturing a host cell transformed with an expression vector containing nucleic acid encoding an IbA A protein, under the appropriate conditions to induce or cause expression of the IbA protein. The conditions appropriate for IbA protein expression will vary with the choice of the expression vector and the host cell, and will be easily ascertained by one skilled in the art through routine experimentation. For example, the use of constitutive promoters in the expression vector will require optimizing the growth and proliferation of the host cell, while the use of an inducible promoter requires the appropriate growth conditions for induction. In addition, in some embodiments, the timing of the harvest is important. For example, the baculoviral systems used in insect cell expression are lytic viruses, and thus harvest time selection can be crucial for product yield.

Appropriate host cells include yeast, bacteria, archebacteria, fungi, and insect and animal cells, including mammalian cells. Of particular interest are *Drosophila melangaster* cells, *Saccharomyces cerevisiae* and other yeasts, *E. coli, Bacillus subtilis*, SF9 cells, C129 cells, 293 cells, Neurospora, BHK, CHO, COS, Pichia Pastoris, etc.

In a preferred embodiment, the IbA proteins are expressed in mammalian cells. Mammalian expression systems are also known in the art, and include retroviral systems. A mammalian promoter is any DNA sequence capable of binding mammalian RNA polymerase and initiating the downstream (3') transcription of a coding sequence for the fusion protein into mRNA. A promoter will have a transcription initiating region, which is usually placed proximal to the 5' end of the coding sequence, and a TATA box, using a located 25–30 base pairs upstream of the transcription initiation site. The TATA box is thought to direct RNA polymerase II to begin RNA synthesis at the correct site. A mammalian promoter will also contain an upstream promoter element (enhancer element), typically located within 100 to 200 base pairs upstream of the TATA box. An upstream promoter element determines the rate at which transcription is initiated and can act in either orientation. Of particular use as mammalian promoters are the promoters from mammalian viral genes, since the viral genes are often highly expressed and have a broad host range. Examples include the SV40 early promoter, mouse mammary tumor virus LTR promoter, adenovirus major late promoter, herpes simplex virus promoter, and the CMV promoter.

Typically, transcription termination and polyadenylation sequences recognized by mammalian cells are regulatory regions located 3' to the translation stop codon and thus, together with the promoter elements, flank the coding sequence. The 3' terminus of the mature mRNA is formed by site-specific post-translational cleavage and polyadenylation. Examples of transcription terminator and polyadenylation signals include those derived form SV40.

The methods of introducing exogenous nucleic acid into mammalian hosts, as well as other hosts, is well known in the art, and will vary with the host cell used. Techniques include dextran-mediated transfection, calcium phosphate precipitation, polybrene mediated transfection, protoplast fusion, electroporation, viral infection, encapsulation of the polynucleotide(s) in liposomes, and direct microinjection of the DNA into nuclei. As outlined herein, a particularly preferred method utilizes retroviral infection, as outlined in PCT US97/01019, incorporated by reference.

As will be appreciated by those in the art, the type of mammalian cells used in the present invention can vary widely. Basically, any mammalian cells may be used, with mouse, rat, primate and human cells being particularly preferred, although as will be appreciated by those in the art, modifications of the system by pseudotyping allows all eukaryotic cells to be used, preferably higher eukaryotes. As is more fully described below, a screen will be set up such that the cells exhibit a selectable phenotype in the presence of a bioactive peptide. As is more fully described below, cell types implicated in a wide variety of disease conditions are particularly useful, so long as a suitable screen may be designed to allow the selection of cells that exhibit an altered phenotype as a consequence of the presence of a peptide within the cell.

Accordingly, suitable cell types include, but are not limited to, tumor cells of all types (particularly melanoma, myeloid leukemia, carcinomas of the lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell) , mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoetic, neural, skin, lung, kidney, liver and myocyte stem cells (for use in screening for differentiation and de-differentiation factors), osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to, Jurkat T cells, NIH3T3 cells, CHO, Cos, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

In one embodiment, the cells may be additionally genetically engineered, that is, contain exogeneous nucleic acid other than the IbA nucleic acid.

In a preferred embodiment, the IbA proteins are expressed in bacterial systems. Bacterial expression systems are well known in the art.

A suitable bacterial promoter is any nucleic acid sequence capable of binding bacterial RNA polymerase and initiating the downstream (3') transcription of the coding sequence of the IbA protein into mRNA. A bacterial promoter has a transcription initiation region which is usually placed proximal to the 5' end of the coding sequence. This transcription initiation region typically includes an RNA polymerase binding site and a transcription initiation site. Sequences encoding metabolic pathway enzymes provide particularly useful promoter sequences. Examples include promoter sequences derived from sugar metabolizing enzymes, such as galactose, lactose and maltose, and sequences derived from biosynthetic enzymes such as tryptophan. Promoters from bacteriophage may also be used and are known in the art. In addition, synthetic promoters and hybrid promoters are also useful; for example, the tac promoter is a hybrid of the trp and lac promoter sequences. Furthermore, a bacterial promoter can include naturally occurring promoters of non-bacterial origin that have the ability to bind bacterial RNA polymerase and initiate transcription.

In addition to a functioning promoter sequence, an efficient ribosome binding site is desirable. In *E. coli*, the ribosome binding site is called the Shine-Delgarno (SD) sequence and includes an initiation codon and a sequence 3-9 nucleotides in length located 3-11 nucleotides upstream of the initiation codon.

The expression vector may also include a signal peptide sequence that provides for secretion of the IbA protein in bacteria. The signal sequence typically encodes a signal peptide comprised of hydrophobic amino acids which direct the secretion of the protein from the cell, as is well known in the art. The protein is either secreted into the growth media (gram-positive bacteria) or into the periplasmic space, located between the inner and outer membrane of the cell (gram-negative bacteria). For expression in bacteria, usually bacterial secretory leader sequences, operably linked to an IbA encoding nucleic acid, are preferred.

The bacterial expression vector may also include a selectable marker gene to allow for the selection of bacterial strains that have been transformed. Suitable selection genes include genes which render the bacteria resistant to drugs such as ampicillin, chloramphenicol, erythromycin, kanamycin, neomycin and tetracycline. Selectable markers also include biosynthetic genes, such as those in the histidine, tryptophan and leucine biosynthetic pathways.

These components are assembled into expression vectors. Expression vectors for bacteria are well known in the art, and include vectors for *Bacillus subtilis, E. coli, Streptococcus cremoris*, and *Streptococcus lividans*, among others.

The bacterial expression vectors are transformed into bacterial host cells using techniques well known in the art, such as calcium chloride treatment, electroporation, and others.

In one embodiment, IbA proteins are produced in insect cells. Expression vectors for the transformation of insect cells, and in particular, baculovirus-based expression vectors, are well known in the art.

In a preferred embodiment, IbA protein is produced in yeast cells. Yeast expression systems are well known in the art, and include expression vectors for *Saccharomyces cerevisiae, Candida albicans* and *C. maltosa, Hansenula polymorpha, Kluyveromyces fragilis* and *K. lactis, Pichia guillerimondii* and *P. pastoris, Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. Preferred promoter sequences for expression in yeast include the inducible GAL1,10 promoter, the promoters from alcohol dehydrogenase, enolase, glucokinase, glucose-6-phosphate isomerase, glyceraldehyde-3-phosphate-dehydrogenase, hexokinase, phosphofructokinase, 3-phosphoglycerate mutase, pyruvate kinase, and the acid phosphatase gene. Yeast selectable markers include ADE2, HIS4, LEU2, TRP1, and ALG7, which confers resistance to tunicamycin; the neomycin phosphotransferase gene, which confers resistance to G418; and the CUP1 gene, which allows yeast to grow in the presence of copper ions.

In addition, the IbA polypeptides of the invention may be further fused to other proteins, if desired, for example to increase expression or stabilize the protein.

In one embodiment, the IbA nucleic acids, proteins and antibodies of the invention are labeled with a label other than the scaffold. By "labeled" herein is meant that a compound has at least one element, isotope or chemical compound attached to enable the detection of the compound. In general, labels fall into three classes: a) isotopic labels, which may be radioactive or heavy isotopes; b) immune labels, which may be antibodies or antigens; and c) colored or fluorescent dyes. The labels may be incorporated into the compound at any position.

Once made, the IbA proteins may be covalently modified. One type of covalent modification includes reacting targeted amino acid residues of an IbA polypeptide with an organic derivatizing agent that is capable of reacting with selected side chains or the N-or C-terminal residues of an IbA polypeptide. Derivatization with bifunctional agents is useful, for instance, for crosslinking an IbA protein to a water-insoluble support matrix or surface for use in the method for purifying anti-IbA antibodies or screening assays, as is more fully described below. Commonly used crosslinking agents include, e.g., 1,1-bis(diazoacetyl)-2-phenylethane, glutaraldehyde, N-hydroxysuccinimide esters, for example, esters with 4-azidosalicylic acid, homobifunctional imidoesters, including disuccinimidyl esters such as 3,3'-dithiobis(succinimidylpropionate), bifunctional maleimides such as bis-N-maleimido-1,8-octane and agents such as methyl-3-[(p-azidophenyl)dithio]propioimidate.

Other modifications include deamidation of glutaminyl and asparaginyl residues to the corresponding glutamyl and aspartyl residues, respectively, hydroxylation of proline and lysine, phosphorylation of hydroxyl groups of seryl or threonyl residues, methylation of the "-amino groups of lysine, arginine, and histidine side chains [T. E. Creighton, Proteins: Structure and Molecular Properties, W. H. Freeman & Co., San Francisco, pp. 79–86 (1983)], acetylaffon of the N-terminal amine, and amidation of any C-terminal carboxyl group.

Another type of covalent modification of the IbA polypepfide included within the scope of this invention comprises altering the native glycosylation pattern of the polypeptide. "Altering the native glycosylation pattern" is intended for purposes herein to mean deleting one or more carbohydrate moieties found in native sequence IbA polypeptide, and/or adding one or more glycosylation sites that are not present in the native sequence IbA polypeptide.

Addition of glycosylation sites to IbA polypeptides may be accomplished by altering the amino acid sequence thereof. The alteration may be made, for example, by the addition of, or substitution by, one or more serine or threonine residues to the native sequence IbA polypeptide (for O-linked glycosylation sites). The IbA amino acid sequence may optionally be altered through changes at the DNA level, particularly by mutating the DNA encoding the IbA polypeptide at preselected bases such that codons are generated that will translate into the desired amino acids.

Another means of increasing the number of carbohydrate moieties on the IbA polypeptide is by chemical or enzymatic coupling of glycosides to the polypeptide. Such methods are described in the art, e.g., in WO87/05330 published Sep. 11, 1987, and in Aplin and Wriston, CRC Crit. Rev. Biochem., pp. 259–306 (1981).

Removal of carbohydrate moieties present on the IbA polypeptide may be accomplished chemically or enzymatically or by mutational substitution of codons encoding for amino acid residues that serve as targets for glycosylabon. Chemical deglycosylation techniques are known in the art and described, for instance, by Hakimuddin, et al., Arch. Biochem. Biophys., 259:52 (1987) and by Edge et al., Anal. Biochem., 118:131 (1981). Enzymatic cleavage of carbohydrate moieties on polypeptides can be achieved by the use of a variety of endo-and exo-glycosidases as described by Thotakura et al., Meth. Enzymol., 138:350 (1987).

Such derivatized moieties may improve the solubility, absorption, permeability across the blood brain barrier, biological half life, and the like. Such moieties or modifications of IbA polypeptides may alternatively eliminate or attenuate any possible undesirable side effect of the protein and the like. Moieties capable of mediating such effects are disclosed, for example, in Remington's Pharmaceutical Sciences, 16th ed., Mack Publishing Co., Easton, Pa. (1980).

Another type of covalent modification of IbA comprises linking the IbA polypeptide to one of a variety of nonproteinaceous polymers, e.g., polyethylene glycol, polypropylene glycol, or polyoxyalkylenes, in the manner set forth in U.S. Pat. Nos. 4,640,835; 4,496,689; 4,301,144; 4,670,417; 4,791,192 or 4,179,337.

IbA polypeptides of the present invention may also be modified in a way to form chimeric molecules comprising an IbA polypeptide fused to another, heterologous polypeptide or amino acid sequence. In one embodiment, such a chimeric molecule comprises a fusion of an IbA polypeptide with a tag polypeptide which provides an epitope to which an anti-tag antibody can selectively bind. The epitope tag is generally placed at the amino-or carboxyl-terminus of the IbA polypeptide. The presence of such epitope-tagged forms of an IbA polypeptide can be detected using an antibody against the tag polypeptide. Also, provision of the epitope tag enables the IbA polypepbde to be readily purified by affinity purification using an anti-tag antibody or another type of affinity matrix that binds to the epitope tag. In an alternative embodiment, the chimeric molecule may comprise a fusion of an IbA polypeptide with an immunoglobulin or a particular region of an immunoglobulin. For a bivalent form of the chimeric molecule, such a fusion could be to the Fc region of an IgG molecule.

Various tag polypeptides and their respective antibodies are well known in the art. Examples include poly-histidine (poly-his) or poly-histidine-glycine (poly-his-gly) tags; the flu HA tag polypeptide and its antibody 12CA5 [Field et al., Mol. Cell. Biol. 8:2159–2165 (1988)]; the c-myc tag and the 8F9, 3C7, 6E10, G4, B7 and 9E10 antibodies thereto [Evan et al., Molecular and Cellular Biology, 5:3610–3616 (1985)]; and the Herpes Simplex virus glycoprotein D (gD) tag and its antibody [Paborsky et al., Protein Engineering, 3(6):547–553 (1990)]. Other tag polypeptdes include the Flag-peptide [Hopp et al., BioTechnology 6:1204–1210 (1988)]; the KT3 epitope peptde [Martin et al., Science 255:192–1944 (1992)]; tubulin epitope peptide [Skinner et al., J. Biol. Chem. 266:15163–15166 (1991)]; and the T7 gene 10 protein peptde tag [Lutz-Freyermuth et al., Proc. Natl. Acad. Sci. U.S.A. 87:6393–6397 (1990)].

In a preferred embodiment, the IbA protein is purified or isolated after expression. IbA proteins may be isolated or purified in a variety of ways known to those skilled in the art depending on what other components are present in the sample. Standard purification methods include electrophoretic, molecular, immunological and chromatographic techniques, including ion exchange, hydrophobic, affinity, and reverse-phase HPLC chromatography, and chromatofocusing. For example, the IbA protein may be purified using a standard anti-library antibody column. Ultrafiltration and diafiltration techniques, in conjunction with protein concentration, are also useful. For general guidance in suitable purification techniques, see Scopes, R., Protein Purification, Springer-Verlag, NY (1982). The degree of purification necessary will vary depending on the use of the IbA protein. In some instances no purification will be necessary.

Once made, the IbA proteins and nucleic acids of the invention find use in a number of applications. In a preferred embodiment, the IbA proteins are administered to a patent to treat an IFN-β-associated disorder.

By "IFN-β associated disorder" or "IFN-β responsive disorder" or "condition" herein is meant a disorder that can be ameliorated by the administration of a pharamaceutical composition comprising an IFN-β or IbA protein, including, but not limited to, multiple sclerosis; idiopathic pulmonary fibrosis; inflammatory diseases; viral diseases; infections caused by papilloma viruses, such as genital warts and condylomata of the uterine cervix; infections caused by hepatitis viruses, such as acute/chronic hepatitis B and non-A, non-B hepatitis (hepatitis C); infections caused by herpes viruses, such as herpes genitalis, herpes zoster, herpes keratitis, and herpes simplex; viral encephalitis; cytomegalovirus pneumonia; prophylaxis of rhinovirus; cancer, including several malignant diseases such as osteosarcoma, basal cell carcinoma, cervical dysplasia, glioma, acute myeloid leukemia, multiple myeloma, Hodgkin's disease, melanoma, renal cancer, liver cancer, and breast cancer.

In a preferred embodiment, a therapeutically effective dose of an IbA protein is administered to a patient in need of treatment. By "therapeutically effective dose" herein is meant a dose that produces the effects for which it is administered. The exact dose will depend on the purpose of the treatment, and will be ascertainable by one skilled in the art using known techniques. In a preferred embodiment, dosages of about 5 μg/kg are used, administered either intraveneously or subcutaneously. As is known in the art, adjustments for IbA protein degradation, systemic versus localized delivery, and rate of new protease synthesis, as well as the age, body weight, general health, sex, diet, time of administration, drug interaction and the severity of the condition may be necessary, and will be ascertainable with routine experimentation by those skilled in the art.

A "patient" for the purposes of the present invention includes both humans and other animals, particularly mammals, and organisms. Thus the methods are applicable to both human therapy and veterinary applications. In the preferred embodiment the patient is a mammal, and in the most preferred embodiment the patient is human.

The term "treatment" in the instant invention is meant to include therapeutic treatment, as well as prophylactic, or suppressive measures for the disease or disorder. Thus, for example, in the case of multiple sclerosis, successful administration of an IbA protein prior to onset of the disease results in "treatment" of the disease. As another example, successful administration of an IbA protein after clinical manifestation of the disease to combat the symptoms of the disease comprises treatment" of the disease. "Treatment" also encompasses administration of an IbA protein after the appearance of the disease in order to eradicate the disease. Successful administration of an agent after onset and after clinical symptoms have developed, with possible abatement of clinical symptoms and perhaps amelioration of the disease, comprises "treatment" of the disease.

Those "in need of treatment" include mammals, in particular humans, already having the disease or disorder, as well as those prone to having the disease or disorder, including those in which the disease or disorder is to be prevented.

In another embodiment, a therapeutically effective dose of an IbA protein, an IbA gene, or an IbA antibody is administered to a patient having a disease involving inappropriate expression of IFN-β. A "disease involving inappropriate expression of a IFN-β" within the scope of the present invention is meant to include diseases or disorders characterized by an overabundance of IFN-β. This overabundance may be due to any cause, including, but not limited to, overexpression at the molecular level, prolonged or accumulated appearance at the site of action, or increased activity of IFN-β relative to normal. Included within this definition are diseases or disorders characterized by a reduction of IFN-β. This reduction may be due to any cause, including, but not limited to, reduced expression at the molecular level, shortened or reduced appearance at the site of action, or decreased activity of IFN-β relative to normal. Such an overabundance or reduction of IFN-β can be measured relative to normal expression, appearance, or activity of IFN-β according to, but not limited to, the assays described and referenced herein.

The administration of the IbA proteins of the present invention, preferably in the form of a sterile aqueous solution, can be done in a variety of ways, including, but not limited to, orally, subcutaneously, intravenously, intranasally, transdermally, intraperitoneally, intramuscularly, intrapulmonary, vaginally, rectally, or intraocularly. In some instances, for example, in the treatment of wounds, inflammation, or multiple sclerosis, the IbA A protein may be directly applied as a solution or spray. Depending upon the manner of introduction, the pharmaceutical composition may be formulated in a variety of ways. The concentration of the therapeutically active IbA protein in the formulation may vary from about 0.1 to 100 weight %. In another preferred embodiment, the concentration of the IbA protein is in the range of 0.003 to 1.0 molar, with dosages from 0.03, 0.05, 0.1, 0.2, and 0.3 millimoles per kilogram of body weight being preferred.

The pharmaceutical compositions of the present invention comprise an IbA protein in a form suitable for administration to a patient. In the preferred embodiment, the pharmaceutical compositions are in a water soluble form, such as being present as pharmaceutically acceptable salts, which is meant to include both acid and base addition salts. "Pharmaceutically acceptable acid addition salt" refers to those salts that retain the biological effectiveness of the free bases and that are not biologically or otherwise undesirable, formed with inorganic acids such as hydrochloric acid, hydrobromic acid, sulfuric acid, nitric acid, phosphoric acid and the like, and organic acids such as acetic acid, propionic acid, glycolic acid, pyruvic acid, oxalic acid, maleic acid, malonic acid, succinic acid, fumaric acid, tartaric acid, citric acid, benzoic acid, cinnamic acid, mandelic acid, methanesulfonic acid, ethanesulfonic acid, p-toluenesulfonic acid, salicylic acid and the like. "Pharmaceutically acceptable base addition salts" include those derived from inorganic bases such as sodium, potassium, lithium, ammonium, calcium, magnesium, iron, zinc, copper, manganese, aluminum salts and the like. Particularly preferred are the ammonium, potassium, sodium, calcium, and magnesium salts. Salts derived from pharmaceutically acceptable organic non-toxic bases include salts of primary, secondary, and tertiary amines, substituted amines including naturally occurring substituted amines, cyclic amines and basic ion exchange resins, such as isopropylamine, trimethylamine, diethylamine, triethylamine, tripropylamine, and ethanolamine.

The pharmaceutical compositions may also include one or more of the following: carrier proteins such as serum albumin; buffers such as NaOAc; fillers such as microcrystalline cellulose, lactose, corn and other starches; binding agents; sweeteners and other flavoring agents; coloring agents; and polyethylene glycol. Additives are well known in the art, and are used in a variety of formulations.

In addition, in one embodiment, the IbA proteins of the present invention are formulated using a process for pharmaceutical compositions of recombinant IFN-β as described in U.S. Pat. No. 5,183,746 which, hereby, is expressly incorporated in its entirety.

In a further embodiment, the IbA proteins are added in a micellular formulation; see U.S. Pat. No. 5,833,948, hereby expressly incorporated by reference in its entirety.

Combinations of pharmaceutical compositions may be administered. Moreover, the compositions may be administered in combination with other therapeutics.

In one embodiment provided herein, antibodies, including but not limited to monoclonal and polyclonal antibodies, are raised against IbA proteins using methods known in the art. In a preferred embodiment, these ant-IbA antibodies are used for immunotherapy. Thus, methods of immunotherapy are provided. By "immunotherapy" is meant treatment of an IFN-β related disorders with an antibody raised against an IbA protein. As used herein, immunotherapy can be passive or active. Passive immunotherapy, as defined herein, is the passive transfer of antibody to a recipient (patient). Active immunization is the induction of antibody and/or T-cell responses in a recipient (patient). Induction of an immune response can be the consequence of providing the recipient with an IbA protein antigen to which antibodies are raised. As appreciated by one of ordinary skill in the art, the IbA protein antigen may be provided by injecting an IbA polypeptide against which antibodies are desired to be raised into a recipient, or contacting the recipient with an IbA protein encoding nucleic acid, capable of expressing the IbA protein antigen, under conditions for expression of the IbA protein antigen.

In another preferred embodiment, a therapeutic compound is conjugated to an antibody, preferably an ant-IbA protein antibody. The therapeutic compound may be a cytotoxic agent. In this method, targeting the cytotoxic agent to tumor tissue or cells, results in a reduction in the number of afflicted cells, thereby reducing symptoms associated with cancer, and IbA protein related disorders. Cytotoxic agents are numerous and varied and include, but are not limited to, cytotoxic drugs or toxins or active fragments of such toxins. Suitable toxins and their corresponding fragments include diptheria A chain, exotoxin A chain, ricin A chain, abrin A chain, curcin, crotin, phenomycin, enomycin and the like. Cytotoxic agents also include radiochemicals made by conjugating radioisotopes to antibodies raised against cell cycle proteins, or binding of a radionuclide to a chelating agent that has been covalently attached to the antibody.

In a preferred embodiment, IbA proteins are administered as therapeutic agents, and can be formulated as outlined above. Similarly, IbA genes (including both the full-length sequence, partial sequences, or regulatory sequences of the IbA coding regions) can be administered in gene therapy applications, as is known in the art. These IbA genes can include antisense applications, either as gene therapy (i.e. for incorporation into the genome) or as antisense compositions, as will be appreciated by those in the art.

In a preferred embodiment, the nucleic acid encoding the IbA proteins may also be used in gene therapy. In gene therapy applications, genes are introduced into cells in order to achieve in vivo synthesis of a therapeutically effective genetic product, for example for replacement of a defective gene. "Gene therapy" includes both conventional gene therapy where a lasting effect is achieved by a single treatment, and the administration of gene therapeutic agents, which involves the one time or repeated administration of a therapeutically effective DNA or mRNA. Antisense RNAs and DNAs can be used as therapeutic agents for blocking the expression of certain genes in vivo. It has already been shown that short anbsense oligonucleotides can be imported into cells where they act as inhibitors, despite their low intracellular concentrations caused by their restricted uptake by the cell membrane. [Zamecnik et al., Proc. Natl. Acad. Sci. U.S.A. 83:4143–4146 (1986)]. The oligonucleotides can be modified to enhance their uptake, e.g. by substituting their negatively charged phosphodiester groups by uncharged groups.

There are a variety of techniques available for introducing nucleic acids into viable cells. The techniques vary depending upon whether the nucleic acid is transferred into cultured cells in vitro, or in vivo in the cells of the intended host. Techniques suitable for the transfer of nucleic acid into mammalian cells in vitro include the use of liposomes, electroporation, microinjection, cell fusion, DEAE-dextran, the calcium phosphate precipitation method, etc. The currently preferred in vivo gene transfer techniques include transfection with viral (typically retroviral) vectors and viral coat protein-liposome mediated transfection [Dzau et al., Trends in Biotechnology 11:205–210 (1993)]. In some situations it is desirable to provide the nucleic acid source with an agent that targets the target cells, such as an antibody specific for a cell surface membrane protein or the target cell, a ligand for a receptor on the target cell, etc. Where liposomes are employed, proteins which bind to a cell surface membrane protein associated with endocytosis may be used for targeting and/or to facilitate uptake, e.g. capsid proteins or fragments thereof tropic for a particular cell type, antibodies for proteins which undergo internalization in cycling, proteins that target intracellular localization and enhance intracellular half-life. The technique of receptor-mediated endocytosis is described, for example, by Wu et al., J. Biol. Chem. 262:4429–4432 (1987); and Wagner et al., Proc. Natl. Acad. Sci. U.S.A. 87:3410–3414 (1990). For review of gene marking and gene therapy protocols see Anderson et al., Science 256:808–813 (1992).

In a preferred embodiment, IbA genes are administered as DNA vaccines, either single genes or combinations of IbA genes. Naked DNA vaccines are generally known in the art. Brower, Nature Biotechnology, 16:1304–1305 (1998). Methods for the use of genes as DNA vaccines are well known to one of ordinary skill in the art, and include placing an IbA gene or portion of an IbA gene under the control of a promoter for expression in a patent in need of treatment. The IbA gene used for DNA vaccines can encode full-length IbA proteins, but more preferably encodes portions of the IbA proteins including peptides derived from the IbA protein. In a preferred embodiment a patient is immunized with a DNA vaccine comprising a plurality of nucleotide sequences derived from an IbA gene. Similarly, it is possible to immunize a patient with a plurality of IbA genes or portions thereof as defined herein. Without being bound by theory, expression of the polypeptide encoded by the DNA vaccine, cytotoxic T-cells, helper T-cells and antibodies are induced which recognize and destroy or eliminate cells expressing IFN-β proteins.

In a preferred embodiment, the DNA vaccines include a gene encoding an adjuvant molecule with the DNA vaccine. Such adjuvant molecules include cytokines that increase the immunogenic response to the IbA polypeptide encoded by the DNA vaccine. Additional or alternative adjuvants are known to those of ordinary skill in the art and find use in the invention.

The following examples serve to more fully describe the manner of using the above-described invention, as well as to set forth the best modes contemplated for carrying out various aspects of the invention. It is understood that these examples in no way serve to limit the true scope of this invention, but rather are presented for illustrative purposes. All references cited herein are incorporated by reference in their entirety.

EXAMPLE 1

DESIGN AND CHARACTERIZATION OF NOVEL IbA PROTEINS BY PDA

Summary: Sequences for novel interferon-beta activity proteins (IbA proteins) were designed by simultaneously optimizing residues in the buried core of the protein using Protein Design Automation (PDA) as described in WO98/47089, U.S. Ser. Nos. 09/058,459, 09/127,926, 60/104,612, 60/158,700, 09/419,351, 60/181,630, 60/186,904, and U.S patent application, entitled *Protein Design Automation For Protein Libraries* (Filed: Apr. 14, 2000; Inventor: Bassil Dahiyat), all of which are expressly incorporated by reference in their entirety. Several core designs were completed, with 20–61 residues considered corresponding to 20^{20}–20^{61} sequence possibilities. Residues unexposed to solvent were designed in order to minimize changes to the molecular surface and to limit the potential for antigenicity of designed novel protein analogues.

Calculations required from 12–19 hours on 16 Silicon Graphics R10000 CPU's. The global optimum sequence from each design was selected for characterization. From 2–11 residues were changed from human IFN-β in the designed proteins, out of 166 residues total.

COMPUTATIONAL PROTOCOLS

Template Structure Preparation:

For this study the crystal structure of human IFN-β as deposited in the PDB data bank was used [PDB record 1AU1; Karpusasetal. Proc. Natl. Acad. Sci. U.S.A. 94(22) :11813–8 (1997)]. Karpasus et al. expressed human IFN-β in CHO cells (glycosylated form) and solved the structure by x-ray crystallography to a resolution of 2.2 Ångstrom. The structure of IFN-β is dimeric containing a zinc ion at the interface and both IFN-β monomers (A-chain and B-chain) are glycosylated at asparagine 80. Although both monomers contained 166 amino acid residues, the coordinates for residues 28 to 30 in the B-monomer were not given in the PDB file 1AU1. PDA calculations were performed for the A-chain and B-chain separately. The zinc ion, all water molecules and the carbohydrate moiety as well as all hydrogen atoms that are present in the PDB file 1AU1 were removed from the structure prior to the PDA calculation.

Design Strategies:

Core residues were selected for design since optimization of these positions can improve stability, although stabilization has been obtained from modifications at other sites as well. Core designs also minimize changes to the molecular surface and thus limit the designed protein's potential for antigenicity. PDA calculations were run on 3 core sequences (see FIG. 3) and in a total of 15 core designs (IFN-β A-chain: Core 1, Core 2, Core 2a, Core 3, Core 4, Core 5, and Core 6; IFN-β B-chain: Core 1, Core 2, Core 2a, Core 3, Core 4, Core 5, Core 6 and Core 7; see below).

PDA Calculations

All PDA calculations were performed with salvation model 2. Solvation model 2 is the solvation model described by Street and Mayo [Fold. Design 3:253–258 (1998)]. If possible, Dead End Elimination (DEE) was run to completion to find the PDA ground state. This was done for the PDA calculations for the A-chain and B-chain of Core 1, Core 2 and Core 2a, as defined below. For the calculation of Core 3, Core 4, Core 5, Core 6 and Core 7, DEE was aborted after the rotamer sequence space was reduced to less than 10^{25} sequences. The DEE calculation was for all the given Core calculation followed by Monte Carlo (MC) minimization and a list of the 1000 lowest energy sequences was generated.

A similar procedure was used for the B-chain, where in a first step the side chain of Lys 33 was minimized for 50 steps followed by an additional 50 steps of minimization of the complete B-chain structure. As the coordinates of residues 28 to 30 are missing in the B-chain, the N-terminus of Cys

31 and the C-terminus of Arg 27 were saturated with a hydrogen atom and the NH_2-group in Cys 31 and the COOH group in Arg 27 were kept fixed during minimization to prevent them from moving too far away from their initial positions.

Before the PDA calculations were started an initial preparation of the structure was performed. For the A-chain, the side chains of Phe 50, Glu 61, Lys 115, Met 117 were minimized with Biograf for 50 steps using conjugate gradient procedure without a Coulomb potential. this is followed by an additional 50 steps of conjugate gradient minimization without a Coulomb potential for the complete structure of the A-chain using Biograf. This minimization procedure was chosen to remove initial bad contacts in the structure.

The PDA calculations for all the designs were run using the a2hl p0 rotamer library. This library is based on the backbone-dependent rotamer library of Dunbrack and Karplus (Dunbrack and Karplus, J. Mol. Biol. 230(2):543–74 (1993); hereby expressly incorporated by reference) but includes more rotamers for the aromatic and hydrophobic amino acids; X_1 and X2 angle values of rotamers for all the aromatic amino acids and X_1 angle values for all the other hydrophobic amino acids were expanded ±1 standard deviation about the mean value reported in the Dunbrack and Karplus library. Typical PDA parameters were used: the van der Waals scale factor was set to 0.9, the H-bond potential well-depth was set to 8.0 kcal/mol, the solvation potential was calculated using type 2 solvation with a nonpolar burial energy of 0.048 kcal/mol and a nonpolar exposure multiplication factor of 1.6, and the secondary structure scale factor was set to 0.0 (secondary structure propensities were not considered). Calculations required from 12–24 hours on 16 Silicon Graphics R10000 CPU's.

Monte Carlo Analysis

Monte Carlo analysis of the sequences produced by PDA shows the ground state (optimal) amino acid and amino acids allowed for each variable position and their frequencies of occurrence (see FIGS. 4 through 29).

EXAMPLE 2

PDA Calculations for the A-chain of IFN-β

Different PDA calculations were performed for the core region of the A-chain of IFN-β. In these calculations the number of positions included in the PDA design were varied and the effect of different PDA parameters on the resulting protein sequences, especially the ground state sequence (SEQ ID NO:4), was analyzed.

A-chain Core 1 Design

By visual inspection, the following residues were identified as belonging to the core of the protein: Leu 6, Gln 10, Asn 14, Cys 17, Leu 21, Ala 55, Ala 56, Thr 58, Ile 59, Met 62, Leu 63, Ile 66, Ile 69, Phe 70, Val 84, Leu 87, Val 91, Gln 94, Leu 98, Ser 118, Leu 122, Tyr 125, Tyr 126, Ile 129, Leu 133, Ala 142, Trp 143, Val 146, Ile 150, Asn 153, Phe 154, Ile 157, and Leu 160. In the first calculation, Cys 17 was not included. Also excluded from the PDA design were Phe 70, Trp 143, and Phe 154, as they are known to be important in the stabilization of the core region, and Gln 10, Thr 58, Gln 94, Ser 118 were excluded as they form side chain H-bonds. Furthermore, residues Tyr 125, Tyr 126 and Asn 153 were not considered as these amino acids are highly conserved in IFN-ps from different organisms as well as Ala 142 as its mutation to Thr is known to lead to loss of function.

Thus, the following positions were included in the PDA design (see also FIG. 3):

6	21	55	56	59	62	63	66	69	84	87	91
Leu	Leu	Ala	Ala	Ile	Met	Leu	Ile	Ile	Val	Leu	Val

98	122	129	133	146	150	157	160
Leu	Leu	Ile	Leu	Val	Ile	Ile	Leu

Met 62 was allowed to change to any PHOBIC amino acid (Ala, Val, Leu, Ile, Phe, Tyr, Trp, Met) and the other residues were allowed to change to Ala, Val, Leu, Ile, Phe, Tyr, Trp and the PDA core solvation potential was used including surface area calculation.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:4):

6	21	55	56	59	62	63	66	69	84	87	91
Leu	Leu	Ala	Ala	Ile	Met	Leu	Ile	Ile	Ile	Phe	Val

98	122	129	133	146	150	157	160
Leu	Leu	Ile	Leu	Val	Ile	Ile	Leu

This sequence shows two mutations from the wild type IFN-β sequence, V84I and L87F (see FIG. 4B) (SEQ ID NO:4).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 4A. Thus, any protein sequence showing mutations at the positions according to FIG. 4A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 4B (SEQ ID NO:4).

A-chain Core 2 Design

To allow more flexibility, all residues that have heavy side chain atoms within a distance of 4 Angstrom of any heavy side chain atom of the amino acids used in the Core 1 calculation were added to the PDA calculation. Thus, Met 1, Gln 10, Asn 14, Cys 17, Phe 38, Phe 50, Thr 58, Glu 61, Phe 70, Glu 81, Gln 94, Ile 95, Leu 102, Lys 115, Tyr 125, Tyr 126, Leu 130, Tyr 138, Thr 144, Arg 147, Leu 151, Asn 153, Phe 154, Arg 159, Thr 161, Tyr 163, and Leu 164 were treated as wild type, such that the conformation of the amino acid side chain could change but not the identity Gln 10, Asn 14, Cys 17, Phe 38, Phe 50, Thr 58, Phe 70, Gln 94, Tyr 125, Tyr 126, Thr 144, Asn 153, Phe 154, Thr 161, and Leu 164 were treated with the PDA core potential for surface area calculation. Ile 95, Leu 102, Arg 147, Leu 151, and Tyr 163 were treated with the PDA boundary potential for surface area calculation. Met 1, Glu 61, Glu 81, Lys 115, Leu 130, Tyr 138, and Arg 159 were treated with the PDA surface potential, but no surface area was calculated.

Thus, the following positions were included in the PDA design (see also FIG. 3):

1	6	10	14	17	21	38	50	55	56	58	59
Met	Leu	Gln	Asn	Cys	Leu	Phe	Phe	Ala	Ala	Thr	Ile

61	62	63	66	69	70	81	84	87	91	94	95
Glu	Met	Leu	Ile	Ile	Phe	Glu	Val	Leu	Val	Gln	Ile

-continued

98	102	115	122	125	126	129	130	133	138	144	146
Leu	Leu	Lys	Leu	Tyr	Tyr	Ile	Leu	Leu	Tyr	Thr	Val

147	150	151	153	154	157	159	160	161	163	164
Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Thr	Tyr	Leu

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:5):

1	6	10	14	17	21	38	50	55	56	58	59
Met	Leu	Gln	Asn	Cys	Leu	Phe	Phe	Ala	Ala	Thr	Ile

61	62	63	66	69	70	81	84	87	91	94	95
Glu	Met	Leu	Ile	Ile	Phe	Glu	Ile	Leu	Ile	Gln	Ile

98	102	115	122	125	126	129	130	133	138	144	146
Phe	Leu	Lys	Ile	Tyr	Tyr	Ile	Leu	Leu	Tyr	Thr	Val

147	150	151	153	154	157	159	160	161	163	164
Arg	Ile	Leu	Asn	Phe	Leu	Arg	Leu	Thr	Tyr	Leu

This sequence shows five mutations from the wild type sequence, V84I, V91I, L98F, L122I, and I157L (see FIG. 5B) (SEQ ID NO:5).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 5A. Thus, any protein sequence showing mutations at the positions according to FIG. 5A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 5B (SEQ ID NO:5).

A-chain Core 2a Design

A calculation similar to Core 2 was performed but now all wild type residues were treated with the PDA core potential including the surface area calculation. This calculation yields the same ground state sequence (SEQ ID NO:5) as resulted from Core 2.

1	6	10	14	17	21	38	50	55	56	58	59
Met	Leu	Gln	Asn	Cys	Leu	Phe	Phe	Ala	Ala	Thr	Ile

61	62	63	66	69	70	81	84	87	91	94	95
Glu	Met	Leu	Ile	Ile	Phe	Glu	Ile	Leu	Ile	Gln	Ile

98	102	115	122	125	126	129	130	133	138	144	146
Phe	Leu	Lys	Ile	Tyr	Tyr	Ile	Leu	Leu	Tyr	Thr	Val

147	150	151	153	154	157	159	160	161	163	164
Arg	Ile	Leu	Asn	Phe	Leu	Arg	Leu	Thr	Tyr	Leu

A-chain Core 3 Design

A slightly larger core region than that used in core 2 was defined. The residues Ser 13, Cys 17, Gly 114, Ser 118, Ala 142, Trp 143, Phe 154, and Thr 161 were added to the PDA design used in core 2a and allowed to change their identity. Ser 13, Ala 142, Trp 143, Phe 154 and Thr 161 could change to any PHOBIC residues except methionine; Cys 17 to any PHOBIC residue plus cysteine, but not to methionine; Gly 114 could become any PHOBIC residue plus glycine, but not methionine; Ser 118 could become any PHOBIC residue plus serine, but no methionine. All these eight were treated with the PDA core potential for surface area calculation. In addition, the following residues were added and treated as wild type using the PDA core potential for surface area calculation: Gln 18, Gln 72, Ser 74, Ser 76, Thr 77, Asn 90, Tyr 132, Lys 136, and Ser 139.

Thus, the following positions were included in the PDA design (see also FIG. 3):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Ser	Asn	Cys	Gln	Leu	Phe	Phe	Ala	Ala

58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr

81	84	87	90	91	94	95	98	102	114	115	118
Glu	Val	Leu	Asn	Val	Gln	Ile	Leu	Leu	Gly	Lys	Ser

122	125	126	129	130	132	133	136	138	139	142	143
Leu	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp

144	146	147	150	151	153	154	157	159	160	161	163
Thr	Val	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Thr	Tyr

164
Leu

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:6):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Phe	Asn	Cys	Gln	Leu	Phe	Phe	Ala	Ala

58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Val	Phe	Gln	Ser	Ser	Thr

81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Phe	Leu	Gly	Lys	Ala

122	125	126	129	130	132	133	136	138	139	142	143
Ile	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp

144	146	147	150	151	153	154	157	159	160	161	163
Thr	Ile	Arg	Ile	Leu	Asn	Phe	Leu	Arg	Leu	Ala	Tyr

164
Leu

This sequence shows 10 mutations from the wild type sequence, S13F, I69V, V84I, V91I, L98F, S118A, L122I, V146I, I157L, and T161A (see FIG. 6B) (SEQ ID NO:6).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 6A. Thus, any protein sequence showing mutations at the positions according to FIG. 6A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. Preferred IbA sequences are shown in FIGS. 6B, 6C, and 6D (SEQ ID NOS: 6–8).

A-chain Core 4 Design

The newly added residues Ser 13, Cys 17, Ser 118, and Thr 161 were now allowed to change to any of the following amino acids: Ala, Val, Leu, Ile, Phe, Tyr, Trp, Asp, Asn, Glu, Gln, Lys, Ser, Thr, His, and Arg, but they were still treated with the PDA core potential for surface area calculation. Otherwise this calculation is identical to Core 3.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:9):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Phe	Asn	Asp	Gln	Leu	Phe	Phe	Ala	Ala
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Val	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Phe	Leu	Gly	Lys	Ala
122	125	126	129	130	132	133	136	138	139	142	143
Ile	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Ile	Arg	Ile	Leu	Asn	Phe	Leu	Arg	Leu	Ala	Tyr
164											
Leu											

This sequence shows 11 mutations from the wild type sequence, S13F, C17D, I69V, V84I, V91I, L98F, S118A, L122I, V146I, I157L, and T161A (see FIG. 7B) (SEQ ID NO:9).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 7A. Thus, any protein sequence showing mutations at the positions according to FIG. 7A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. Preferred IbA sequences are shown in FIGS. 7B, 7C, and 7D (SEQ ID NOS:9–11).

A-chain Core 5 Design

A slightly different change in the identities of the amino acids than in Core4 calculation was now allowed. Leu 6, Leu 21, Ala 55, Ala 56, Ile 59, Leu 63, Ile 66, Ile 69, Val 84, Val 91, Leu 122, Ile 129, Leu 133, Ala 142, Trp 143, Val 146, Ile 150, Phe 154, Ile 157, and Leu 160 could change to any PHOBIC residue except methionine. Met 62 was allowed to change to any PHOBIC amino acid residue; Leu 87, Leu98, and Gly 114 were allowed to change to Ala, Val, Leu, Ile, Gly; and Ser 13, Cys 17, Ser 118, and Thr 161 could change to Ala, Gly, Ser, Thr, Glu, Asp, Gln, Asn, or Cys. All the other residues were treated as wild type as was done in the Core 4 calculation.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:12):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Glu	Asn	Asp	Gln	Leu	Phe	Phe	Ala	Ala
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Gly	Lys	Ser
122	125	126	129	130	132	133	136	138	139	142	143
Leu	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Ile	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Cys	Tyr
164											
Leu											

This sequence shows 7 mutations from the wild type sequence, S13E, C17D, V84I, V91I, S118C, V146I, and T161C (see FIG. 8B) (SEQ ID NO:12).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 8A. Thus, any protein sequence showing mutations at the positions according to FIG. 8A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. Preferred IbA sequences are shown in FIGS. 8B, 8C, and 8D (SEQ ID NOS:12–14). A DNA library can be generated to mirror the probability table of FIG. 8A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

A-chain Core 6 Design

A similar calculation to Core 5 was performed where now at positions 13, 17, 113, and 117 no cysteine was allowed to occur.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:15):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Glu	Asn	Asp	Gln	Leu	Phe	Phe	Ala	Ala
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Val	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Gly	Lys	Asn
122	125	126	129	130	132	133	136	138	139	142	143
Ile	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Ile	Arg	Ile	Leu	Asn	Phe	Leu	Arg	Leu	Ala	Tyr
164											
Leu											

This sequence shows 10 mutations from the wild type sequence, S13E, C17D, 169V, V84I, V91I, S118A, L122I, V146I, I157L, and T161A (see FIG. 9B) (SEQ ID NO:15).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 9A. Thus, any protein sequence showing mutations at the positions according to FIG. 9A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. Preferred IbA sequences are shown in FIGS. 9B, 9C, and 9D (SEQ ID NOS:15–17). A DNA library can be generated to mirror the probability table of FIG. 9A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

EXAMPLE 3

PDA Calculations for the B-chain of IFN-β

For the B-chain, PDA calculations similar to those of the A-chain were performed.

B-chain Core 1 Design

The same positions as for the A-chain Core 1 calculation were used in the PDA design for the B-chain: Leu 6, Leu 21, Ala 55, Ala 56, Ile 59, Met 62, Leu 63, Ile 66, Ile 69, Val 84, Leu 98, Leu 122, Ile 129, Leu 133, Val 146, Ile 150, Ile 157, and Leu 160.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:1 8):

6	21	55	56	59	62	63	66	69	84	87	91
Leu	Leu	Ala	Ala	IIe	Met	Leu	Ile	Ile	Ile	Phe	Val

98	122	129	133	146	150	157	160
Leu	Leu	Ile	Leu	Val	Ile	Ile	Leu

This sequence shows two mutations from the wild type IFN-β sequence, V84I and L87F, and is identical with the ground state sequence generated for the A-chain (see FIG. 10B) (SEQ ID NO:18).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 10A. Thus, any protein sequence showing mutations at the positions according to FIG. 10A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 10B (SEQ ID NO:18). A DNA library can be generated to mirror the probability table of FIG. 10A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 2 Design

A calculation similar to that for the A-chain Core 2 design was performed for the B-chain.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:19):

1	6	10	14	17	21	38	50	55	56	58	59
Met	Leu	Gln	Asn	Cys	Leu	Phe	Phe	Ala	Leu	Thr	Ile

61	62	63	66	69	70	81	84	87	91	94	95
Glu	Met	Phe	Ile	Ile	Phe	Glu	Ile	Phe	Ile	Gln	Ile

98	102	115	122	125	126	129	130	133	138	144	146
Leu	Leu	Lys	Phe	Tyr	Tyr	Ile	Leu	Leu	Tyr	Thr	Val

147	150	151	153	154	157	159	160	161	163	164
Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Thr	Tyr	Leu

This sequence shows six mutations from the wild type sequence, A56L, L63F, V84I, L87F, V91I, and L122F (see FIG. 11B) (SEQ ID NO:19).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 11A. Thus, any protein sequence showing mutations at the positions according to FIG. 11A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 11B (SEQ ID NO:19). A DNA library can be generated to mirror the probability table of FIG. 11A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 2a Design

A calculation similar to that for the A-chain Core 2a design was performed for the B-chain. This calculation yields the same ground state sequence (SEQ ID NO:i9) as resulted from Core 2.

1	6	10	14	17	21	38	50	55	56	58	59
Met	Leu	Gln	Asn	Cys	Leu	Phe	Phe	Ala	Ala	Thr	Ile

61	62	63	66	69	70	81	84	87	91	94	95
Glu	Met	Leu	Ile	Ile	Phe	Glu	Ile	Phe	Ile	Gln	Ile

98	102	115	122	125	126	129	130	133	138	144	146
Leu	Leu	Lys	Phe	Tyr	Tyr	Ile	Leu	Leu	Tyr	Thr	Val

147	150	151	153	154	157	159	160	161	163	164
Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Thr	Tyr	Leu

This sequence shows six mutations from the wild type sequence, A56L, L63F, V84I, L87F, V91I, and L122F.

B-chain Core 3 Design

A calculation similar to that for the A-chain Core 3 was performed for the B-chain, but instead of residue Gln 18, Phe15 was included in the wild type PDA residue list.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:20):

1	6	10	13	14	15	17	21	38	50	55	56
Met	Leu	Gln	Leu	Asn	Phe	Cys	Leu	Phe	Phe	Ala	Leu

58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr

81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Phe	Lys	Leu

122	125	126	129	130	132	133	136	138	139	142	143
Ile	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp

144	146	147	150	151	153	154	157	159	160	161	163
Thr	Val	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Ala	Tyr

164
Leu

This sequence shows 8 mutations from the wild type sequence, S13L, A56L, V84I, V91I, G114F, S118L, L122I, and T161A (see FIG. 12B) (SEQ ID NO:20).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 12A. Thus, any protein sequence showing mutations at the positions according to FIG. 12A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 12B (SEQ ID NO:20). A DNA library can be generated to mirror the probability table of FIG. 12A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 4 Design

A calculation similar to that for the A-chain Core 4 design was performed for the B-chain, but instead of residue Gln 18, Phe 15 was included in the wild type PDA residue list.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:21):

1	6	10	13	14	15	17	21	38	50	55	56
Met	Leu	Gln	Leu	Asn	Phe	Ala	Leu	Phe	Phe	Ala	Leu
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Phe	Asn	Leu	Gln	Ile	Leu	Leu	Phe	Lys	Leu
122	125	126	129	130	132	133	136	138	139	142	143
Ile	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Val	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Glu	Tyr
164											
Leu											

This sequence shows 10 mutations from the wild type sequence, S13L, C17A, A56L, V84I, L87F, V91L, G114F, S118L, L122I, and T161E (see FIG. 13B) (SEQ ID NO:21).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 13A. Thus, any protein sequence showing mutations at the positions according to FIG. 13A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 13B (SEQ ID NO:21). A DNA library can be generated to mirror the probability table of FIG. 13A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 5 Design

A calculation similar to that for the A-chain Core 5 design was performed for the B-chain. Now, Gln 18 was included in the wild type PDA residue list, exactly as was done in the Core 5 calculation for the A-chain.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:22):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Glu	Asn	Cys	Gln	Leu	Phe	Phe	Ala	Leu
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Leu	Lys	Glu
122	125	126	129	130	132	133	136	138	139	142	143
Leu	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Val	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Glu	Tyr
164											
Leu											

This sequence shows 7 mutations from the wild type sequence, S13E, A56L, V84I, V91I, G114L, S118E, and T161E (see FIG. 14B) (SEQ ID NO:22).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 14A. Thus, any protein sequence showing mutations at the positions according to FIG. 14A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 14B (SEQ ID NO:22). A DNA library can be generated to mirror the probability table of FIG. 14A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 6 Design

A similar calculation similar to that for the A-chain Core 6 design was performed for the B-chain.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:23):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Ser	Asn	Thr	Gln	Leu	Phe	Phe	Ala	Leu
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Leu	Lys	Glu
122	125	126	129	130	132	133	136	138	139	142	143
Leu	Tyr	Tyr	Ile	Leu	Tyr	Leu	Lys	Tyr	Ser	Ala	Trp
144	146	147	150	151	153	154	157	159	160	161	163
Thr	Val	Arg	Ile	Leu	Asn	Phe	Ile	Arg	Leu	Glu	Tyr
164											
Leu											

This sequence shows 7 mutations from the wild type sequence, C17T, A56L, V84I, V91I, G114L, S118E, and T161E (see FIG. 15B) (SEQ ID NO:23).

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 15A. Thus, any protein sequence showing mutations at the positions according to FIG. 15A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 15B (SEQ ID NO:23). A DNA library can be generated to mirror the probability table of FIG. 15A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

B-chain Core 7 Design

A similar calculation similar to that of the B-chain Core 6 design was performed. Now Gly 114 is treated as a wild type residue.

The PDA calculation resulted in the following ground state sequence (SEQ ID NO:24):

1	6	10	13	14	17	18	21	38	50	55	56
Met	Leu	Gln	Ser	Asn	Thr	Gln	Leu	Phe	Phe	Ala	Leu
58	59	61	62	63	66	69	70	72	74	76	77
Thr	Ile	Glu	Met	Leu	Ile	Ile	Phe	Gln	Ser	Ser	Thr
81	84	87	90	91	94	95	98	102	114	115	118
Glu	Ile	Leu	Asn	Ile	Gln	Ile	Leu	Leu	Gly	Lys	Glu
122	125	126	129	130	132	133	136	138	139	142	143

-continued

```
Leu Tyr Tyr Ile Leu Tyr Leu Lys Tyr Ser Ala Trp

144 146 147 150 151 153 154 157 159 160 161 163
Thr Val Arg Ile Leu Asn Phe Ile Arg Leu Glu Tyr

164
Leu
```

This sequence shows 6 mutations from the wild type sequence, C17T, A56L, V84I, V91I, S118E, and T161E (see FIG. 16B) (SEQ ID NO:24). With the exception of position 114, now remaining glycine, the ground state sequence is identical to that of Core 6 for the B-chain.

Using Monte Carlo technique a list of low energy sequences was generated. The analysis of the lowest 1000 protein sequences generated by Monte Carlo leads to the mutation pattern shown in FIG. 16A. Thus, any protein sequence showing mutations at the positions according to FIG. 16A will potentially generate a more stable and active IbA. In particular those protein sequences found among the list of the lowest 101 MC generated sequences (data not shown) have a high potential to result in a more stable and active IbA. A preferred IbA sequence is shown in FIG. 16B (SEQ ID NO:24). A DNA library can be generated to mirror the probability table of FIG. 16A that comprises at least one sequence that is more stable and/or active than wild type IFN-β.

SEQUENCE LISTING

```
<160> NUMBER OF SEQ ID NOS: 24

<210> SEQ ID NO 1
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 1

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Val Glu Asn Leu Leu Ala Asn Val Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ser Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165

<210> SEQ ID NO 2
<211> LENGTH: 757
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 2

atgaccaaca agtgtctcct ccaaattgct ctcctgttgt gcttctccac tacagctctt      60

tccatgagct acaacttgct tggattccta caaagaagca gcaattttca gtgtcagaag     120

ctcctgtggc aattgaatgg gaggcttgaa tattgcctca aggacaggat gaactttgac     180
```

-continued

```
atccctgagg agattaagca gctgcagcag ttccagaagg aggacgccgc attgaccatc   240

tatgagatgc tccagaacat ctttgctatt ttcagacaag attcatctag cactggctgg   300

aatgagacta ttgttgagaa cctcctggct aatgtctatc atcagataaa ccatctgaag   360

acagtcctgg aagaaaaact ggagaaagaa gattttacca ggggaaaact catgagcagt   420

ctgcacctga aaagatatta tgggaggatt ctgcattacc tgaaggccaa ggagtacagt   480

cactgtgcct ggaccatagt cagagtggaa atcctaagga acttttactt cattaacaga   540

cttacaggtt acctccgaaa ctgaagatct cctagcctgt ccctctggga ctggacaatt   600

gcttcaagca ttcttcaacc agcagatgct gtttaagtga ctgatggcta atgtactgca   660

aatgaaagga cactagaaga ttttgaaatt tttattaaat tatgagttat ttttatttat   720

ttaaatttta ttttggaaaa taaattattt ttggtgc                            757


<210> SEQ ID NO 3
<211> LENGTH: 21
<212> TYPE: PRT
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 3

Met Thr Asn Lys Cys Leu Leu Gln Ile Ala Leu Leu Leu Cys Phe Ser
1               5                   10                  15

Thr Thr Ala Leu Ser
            20


<210> SEQ ID NO 4
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 4

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Phe Leu Ala Asn Val Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ser Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 5
```

-continued

```
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 5

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ser Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 6
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 6

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Phe Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Val Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
```

```
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 7
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 7

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Tyr Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Val Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Val Ser Leu His Val Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 8
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 8

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Phe Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110
```

-continued

```
Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 9
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 9

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Phe Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Val Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 10
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 10

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Tyr Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Val Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80
```

-continued

```
Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Val Ser Leu His Val Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 11
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 11

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Phe Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Phe Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 12
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 12

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
```

-continued

```
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Cys Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Cys Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 13
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 13

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ala Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Cys Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Cys Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 14
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 14
```

-continued

```
Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Cys Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Cys Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 15
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 15

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Val Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 16
<211> LENGTH: 166
```

```
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 16

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Ile Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 17
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 17

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Asp Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ala Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Leu Asn Arg Leu
145                 150                 155                 160
```

-continued

```
Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 18
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 18

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Ala Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Phe Leu Ala Asn Val Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ser Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 19
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 19

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Phe Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Phe Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Ser Ser Leu His Phe Lys Arg Tyr Tyr Gly Arg
```

-continued

```
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Thr Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 20
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 20

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Leu Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Phe Lys Leu Met Leu Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Ala Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 21
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 21

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Leu Asn Phe Gln
1               5                   10                  15

Ala Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80
```

-continued

```
Glu Thr Ile Ile Glu Asn Phe Leu Ala Asn Leu Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Phe Lys Leu Met Leu Ser Leu His Ile Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Glu Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 22
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 22

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Glu Asn Phe Gln
1               5                   10                  15

Cys Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Leu Lys Leu Met Glu Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Glu Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 23
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 23

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Thr Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45
```

-continued

```
Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Leu Lys Leu Met Glu Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Glu Gly Tyr Leu Arg Asn
                165


<210> SEQ ID NO 24
<211> LENGTH: 166
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic

<400> SEQUENCE: 24

Met Ser Tyr Asn Leu Leu Gly Phe Leu Gln Arg Ser Ser Asn Phe Gln
1               5                   10                  15

Thr Gln Lys Leu Leu Trp Gln Leu Asn Gly Arg Leu Glu Tyr Cys Leu
            20                  25                  30

Lys Asp Arg Met Asn Phe Asp Ile Pro Glu Glu Ile Lys Gln Leu Gln
        35                  40                  45

Gln Phe Gln Lys Glu Asp Ala Leu Leu Thr Ile Tyr Glu Met Leu Gln
    50                  55                  60

Asn Ile Phe Ala Ile Phe Arg Gln Asp Ser Ser Ser Thr Gly Trp Asn
65                  70                  75                  80

Glu Thr Ile Ile Glu Asn Leu Leu Ala Asn Ile Tyr His Gln Ile Asn
                85                  90                  95

His Leu Lys Thr Val Leu Glu Glu Lys Leu Glu Lys Glu Asp Phe Thr
            100                 105                 110

Arg Gly Lys Leu Met Glu Ser Leu His Leu Lys Arg Tyr Tyr Gly Arg
        115                 120                 125

Ile Leu His Tyr Leu Lys Ala Lys Glu Tyr Ser His Cys Ala Trp Thr
    130                 135                 140

Ile Val Arg Val Glu Ile Leu Arg Asn Phe Tyr Phe Ile Asn Arg Leu
145                 150                 155                 160

Glu Gly Tyr Leu Arg Asn
                165
```

I claim:

1. A non-naturally occurring interferon-beta activity (IbA) protein comprising at least fifteen amino acid substitutions as compared to human IFN-β protein (SEQ ID NO: 1), wherein said substitutions are selected from amino acid residues at positions 6, 13, 17, 21, 56, 59, 61, 62, 63, 66, 69, 84, 87, 91, 98, 102, 114, 118, 122, 129, 146, 150, 154, 157, 160, and 161, wherein said protein exhibits at least 50% of the biological activity of human IFN-β protein.

2. The non-naturally occurring IbA protein according to claim 1, wherein said amino acid substitutions are selected from positions 13, 17, 56, 63, 69, 84, 87, 91, 98, 114, 118, 122, 146, 157, and 161.

3. The non-naturally occurring IbA protein according to claim 2, wherein said substitutions are selected from the group of substitutions consisting of S13F, S13Y, S13E, S13A, S13L, C17D, C17A, C17T, A56L, L63F, I69V, V84I, V91I, L98F, G114F, G114L, S118L, S118E, S118A, S118V,

S118C, L122I, L122F, L122V, V146I, I157L, T161A, T161E, and T161C.

4. The non-naturally occurring IbA protein according to claim 1 comprising substitutions at positions 13, 17, 69, 84, 87, 91, 98, 118, 122, 146, 157, and 161.

5. The non-naturally occurring IbA protein according to claim 4, wherein said substitutions are selected from the group of substitutions consisting of S13F, S13Y, S13E, S13A, C17D, 169V, V84I, L87F, V91I, L98F, S118A, S118V, S118C, L122I, L122F, I157L, T161A, and T161C.

6. The non-naturally occurring IbA protein according to claim 1 comprising substitutions at positions 13, 17, 56, 63, 84, 87, 91, 114, 118, 122, and 161.

7. The non-naturally occurring IbA protein according to claim 6, wherein said substitutions are selected from the group of substitutions consisting of S13E, S13L, C17A, C17T, A56L, L63F, V84I, L87F, V91I, G114F, G114L, S118L, S118E, L122I, L122F, T161A, and T161E.

8. A pharmaceutical composition comprising an IbA protein according to claim 1 and a pharmaceutical carrier.

9. A non-naturally occurring protein according to claim 1 wherein said biological activity is the ability to bind to an IFN receptor.

10. A non-naturally occurring protein according to claim 1 wherein said biological activity is the ability to inhibit cell proliferation.

11. A non-naturally occurring protein according to claim 1 wherein said biological activity is the ability to inhibit viral infections.

12. A recombinant nucleic acid encoding the non-naturally occurring IbA protein of claim 1.

13. An expression vector comprising the recombinant nucleic acid of claim 12.

14. A host cell comprising the recombinant nucleic acid of claim 12.

15. A host cell comprising the expression vector of claim 13.

16. A method of producing a non-naturally occurring IbA protein comprising culturing the host cell of claim 15 under conditions suitable for expression of said nucleic acid.

17. The method according to claim 16 further comprising recovering said IbA protein.

* * * * *